

Henning Lobin, Roman Schneider, Andreas Witt (Hrsg.)

DIGITALE INFRA-STRUKTUREN FÜR DIE GERMANISTISCHE FORSCHUNG

GERMANISTISCHE SPRACHWISSENSCHAFT UM 2020





Henning Lobin, Roman Schneider und Andreas Witt (Hrsg.)
Digitale Infrastrukturen für die germanistische Forschung

Germanistische Sprachwissenschaft um 2020

Herausgegeben von
Albrecht Plewnia und Andreas Witt

Band 6



Digitale Infrastrukturen für die germanistische Forschung

Herausgegeben von
Henning Lobin, Roman Schneider und Andreas Witt

DE GRUYTER

Die Open-Access-Publikation dieses Bandes wurde gefördert vom Institut für Deutsche Sprache, Mannheim.

ISBN 978-3-11-053675-1
e-ISBN (PDF) 978-3-11-053866-3
e-ISBN (EPUB) 978-3-11-053681-2



Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

publiziert von Walter de Gruyter GmbH, Berlin/Boston
Foto Einbandabbildung: © Oliver Schonefeld, Institut für Deutsche Sprache, Mannheim
Portrait Ludwig M. Eichinger, Seite V: © David Ausserhofer, Leibniz-Gemeinschaft
Satz: Meta Systems Publishing & Printservices GmbH, Wustermark
Druck und Bindung: CPI books GmbH, Leck

www.degruyter.com



Ludwig M. Eichinger gewidmet

Vorwort

Wo steht die germanistische Sprachwissenschaft aktuell? Der vorliegende Band mit dem Titel „Digitale Infrastrukturen für die germanistische Forschung“ ist der sechste Teil einer auf sechs Bände angelegten Reihe, die eine zwar nicht exhaustive, aber doch umfassende Bestandsaufnahme derjenigen Themenfelder innerhalb der germanistischen Linguistik bieten will, die im Kontext der Arbeiten des Instituts für Deutsche Sprache in den letzten Jahren für das Fach von Bedeutung waren und in den kommenden Jahren von Bedeutung sein werden (und von denen nicht wenige auch vom Institut für Deutsche Sprache bedient wurden und werden). Jeder einzelne Band behandelt ein abgeschlossenes Themengebiet und steht insofern für sich; in der Zusammenschau aller Bände ergibt sich ein Panorama der „Germanistischen Sprachwissenschaft um 2020“.

Anlass des Erscheinens dieser Bände ist der Eintritt des langjährigen Direktors des Instituts für Deutsche Sprache, Ludwig M. Eichinger, in den Ruhestand. Ludwig M. Eichinger leitete das Institut von 2002 bis 2018. Seine akademische Laufbahn begann er als Wissenschaftlicher Assistent an der Universität Bayreuth; anschließend war er Heisenberg-Stipendiat an der Ludwig-Maximilians-Universität München. Ab 1990 hatte er eine Fiebiger-Professur für Deutsche Sprachwissenschaft an der Universität Passau inne, 1997 wurde er auf den Lehrstuhl für Deutsche Philologie an der Christian-Albrechts-Universität zu Kiel berufen. Mit seiner Ernennung zum Direktor des Instituts für Deutsche Sprache im Jahr 2002 wurde er auch Ordinarius für Germanistische Linguistik an der Universität Mannheim. Ludwig M. Eichinger ist Ehrendoktor der Pannonischen Universität Veszprém und der Universität Bukarest. Er ist Mitglied der Akademie der Wissenschaften und der Literatur zu Mainz und der Österreichischen Akademie der Wissenschaften; außerdem ist er Ständiger Gastprofessor an der Beijing Foreign Studies University.

Ludwig M. Eichinger hat das Institut in den Jahren seines Wirkens entscheidend geprägt; in Anerkennung und Dankbarkeit seien ihm diese Bände gewidmet.

Albrecht Plewnia und Andreas Witt
– Reihenherausgeber –

https://doi.org/10.1515/9783110538663-203

Inhalt

Henning Lobin, Roman Schneider und Andreas Witt

Organisierte Kooperativität – Forschungsinfrastrukturen für die germanistische Linguistik

Abstract: Der vorliegende Band befasst sich mit dem Stand und der Entwicklung von Forschungsinfrastrukturen für die germanistische Linguistik und einigen angrenzenden Bereichen. Einen zentralen Aspekt dabei bildet die Notwendigkeit, Kooperativität in der Wissenschaft im institutionellen Sinne, aber auch in Hinsicht auf die wissenschaftliche Praxis zu organisieren. Dies geschieht in Verbünden als Kooperationsstrukturen, wobei Sprachwissenschaft und Sprachtechnologie miteinander verbunden werden. Als zentraler Forschungsressource kommen dabei Korpora und ihrer Erschließung durch spezielle, linguistisch motivierte Informationssysteme besondere Bedeutung zu. Auf der Ebene der Daten werden durch Annotations- und Modellierungsstandards die Voraussetzung für eine nachhaltige Nutzbarkeit derartiger Ressourcen geschaffen.

Keywords: Kooperation, Forschungsverbund, Infrastruktur, Sprachwissenschaft, Sprachtechnologie, Korpus, Informationssystem, Annotation, Modellierung

1 Einführung

Noch vor wenigen Jahren wäre ein Band wie der vorliegende zu digitalen Infrastrukturen für die sprachgermanistische Forschung kaum zu realisieren gewesen. Das liegt nicht allein daran, dass die Digitalisierung erst seit etwa 20 Jahren nach und nach ihre volle Wucht auch in den Geisteswissenschaften entfaltet hat. Forschungsinfrastrukturen lassen sich nicht ohne Kooperation

Henning Lobin, Justus-Liebig-Universität, Institut für Germanistik, Otto-Behaghel-Str. 10 D, D-35394 Gießen, E-Mail: Henning.Lobin@germanistik.uni-giessen.de
Roman Schneider, Institut für Deutsche Sprache, R5 6–13, D-68161 Mannheim, E-Mail: schneider@ids-mannheim.de
Andreas Witt, Universität zu Köln, Institut für Digital Humanities / Sprachliche Informationsverarbeitung & Institut für Deutsche Sprache, Mannheim, E-Mail: andreas.witt@uni-koeln.de & witt@ids-mannheim.de

https://doi.org/10.1515/9783110538663-001

zwischen den Wissenschaftlerinnen und Wissenschaftlern entwickeln und betreiben, und das Prinzip der Kooperativität war in der geisteswissenschaftlichen Forschung nicht so ausgeprägt wie in Disziplinen, die schon immer auf Großgeräte angewiesen waren. Zum „Großgerät“ der germanistischen Linguistik sind heute vernetzte (Korpus-)Infrastruktursysteme geworden, und dieser Band will Stand und Perspektiven dieses neuen, wichtigen Bereichs behandeln. Beginnend mit den Strukturen der Kooperation in Verbünden wird der Gegenstand theoretisch, methodisch und beispielhaft empirisch entfaltet. Fallstudien, wie Sprachkorpora in Verbindung mit sprachtechnologischen Verfahren zur Erkenntnisgewinnung eingesetzt werden, die hypermediale Vermittlung derart erarbeiteter Forschungsergebnisse sowie exemplarische Korpussysteme vermitteln ein Bild von den Möglichkeiten, die aufgrund von Forschungsinfrastrukturen schon heute bestehen. Eine zentrale Grundlage dafür spielen Verabredungen zur Anreicherung von Texten mit Metadaten und wiederkehrenden Datenstrukturen. All diese Aspekte werden im Folgenden in vier Kapiteln behandelt.

Ludwig M. Eichinger hat in den 16 Jahren seiner Tätigkeit als Direktor des Instituts für Deutsche Sprache (IDS) in Mannheim die Bedeutung dieser Entwicklungen so frühzeitig erkannt, dass das IDS nicht nur für die germanistische Linguistik, sondern für die Sprachwissenschaft in Deutschland überhaupt in vielen Bereichen zu einem Zentrum der Infrastrukturentwicklung werden konnte. Die Beiträge in diesem Band zeigen, dass das IDS aufgrund dieser Weichenstellung heute nicht nur in institutioneller Hinsicht, sondern auch bei Sprachressourcen und in der korpuslinguistischen Forschung eine zentrale Position in der Forschungslandschaft einnimmt.

2 Zu den Beiträgen in diesem Band

2.1 Kapitel I – Kooperationen und Verbünde

Im ersten Teil des Bandes, „Kooperationen und Verbünde“, wird in vier Beiträgen die gegenwärtige Situation im Bereich von Forschungsinfrastrukturen und -ressourcen beleuchtet. Im einleitenden Beitrag legt Thomas **Gloning** dar, auf welcher Traditionsgrundlage in einem Fach wie der Germanistik die heutige Entwicklung von Forschungsinfrastukturen zu betrachten ist und wie sehr auch bislang schon Formen der Kollaboration den wissenschaftlichen Diskurs geprägt haben. Trotzdem führt die Digitalisierung auch in dieser Disziplin zu massiven Veränderungen, die eine Neubestimmung zukünftiger Aufgaben in Funktionsbereichen wie Kommunikation, Information und Publikationswesen

als notwendig erscheinen lässt. Glonings Beitrag mündet in die Formulierung von sechs Aufgabenbereichen für den Ausbau von Infrastrukturangeboten aus der Perspektive wissenschaftlicher Nutzer.

Auch Erhard **Hinrichs** stellt die aktuelle Entwicklung von Infrastrukturen für Forschungsdaten in einen historischen Kontext: In der Entwicklung der Sprachwissenschaft im 20. Jahrhundert ist schon lange die Tendenz zu einer Verbreiterung ihrer empirischen Grundlagen zu verzeichnen. Mit der Digitalisierung treten dabei nicht nur viel mehr, sondern auch andere Arten von Sprachdaten in Erscheinung, und durch diese werden besondere Anforderungen an Forschungsinfrastrukturen gestellt. Hinrichs exemplifiziert anhand des Verbundprojekts CLARIN, wie solchen Anforderungen in internationalen Verbünden begegnet werden kann und dabei vielfältige Rückwirkungen auf nationale Planungen zu verzeichnen sind.

Stefan **Schmunk**, Frank **Fischer**, Mirjam **Blümm** und Wolfram **Horstmann** setzen in ihrem Beitrag sogar noch einen Schritt früher an: Sie stellen die Entwicklung geistes- und sozialwissenschaftlicher Forschungsinfrastrukturen insgesamt dar, da diese in vielen Disziplinen ausgehend von den existierenden Infrastruktureinrichtungen wie wissenschaftlichen Bibliotheken bereits seit den 1970er Jahren zunehmend zum Thema geworden sind. Die Digitalisierung bedeutet dabei nicht nur eine Chance, sondern produziert selbst auch neue Probleme wie die nachträgliche digitale Erfassung analoger Datenträger. Ähnlich wie im Bereich der Sprachwissenschaft mit CLARIN existiert für die Geistes- und Sozialwissenschaften insgesamt ein internationaler Forschungsverbund, DARIAH, der eine nationale Spiegelung in Deutschland erfahren hat. Schmunk et al. lassen die Darstellung von DARIAH in die Formulierung von Designprinzipien münden, die bei der Entwicklung digitaler Forschungsinfrastrukturen zu beachten sind.

Karlheinz **Mörth** und Tanja **Wissik** wenden den Blick in ein anderes deutschsprachiges Land. Sie zeigen, wie in Österreich in verschiedenen Schwerpunktbereichen Sprachressourcen aufgebaut worden sind. Anders als in Deutschland besitzt Österreich mit dem Austrian Centre for Digital Humanities (ACDH) an der Österreichischen Akademie der Wissenschaft einen zentralen Knotenpunkt für eine Vielzahl forschungsinfrastruktureller Aktivitäten, der auch als österreichischer Partner sowohl im CLARIN- als auch im DARIAH-Netzwerk fungiert.

2.2 Kapitel II – Sprachwissenschaft und Sprachtechnologie

Der zweite Teil des vorliegenden Bandes, „Sprachwissenschaft und Sprachtechnologie“, befasst sich mit der Nutzung sprachwissenschaftlicher Forschungs-

infrastruktur bei der Beantwortung konkreter Forschungsfragen. Hannah **Kermes** und Elke **Teich** entwickeln in ihrem Beitrag eine generische Methodik für die Erstellung und Analyse von Textkorpora, die bei den Rohdaten ansetzt und über Vorverarbeitung und linguistische Annotation unter Verwendung automatisierter Verfahren zu einer standardisierten Grundlage für empirische Analysen führt. Wie darauf basierende Korpusanalysen durchgeführt werden können, erläutern sie an einem Beispiel, das insbesondere das Wechselspiel zwischen den vorgegebenen Möglichkeiten derartiger Infrastruktursysteme und stets notwendigen individuellen Anpassungen und Ergänzungen in den Blick nimmt.

Auch Kerstin **Eckart**, Markus **Gärtner**, Jonas **Kuhn** und Katrin **Schweitzer** befassen sich in ihrem Beitrag mit methodischen Aspekten, hier allerdings bezogen auf Korpora gesprochener Sprache. Einen zentralen Aspekt ihrer Überlegungen bilden Qualität und Konsistenz der Korpusdaten, für die sie als einen praktikablen Kompromiss die „Silberstandard-Methode“ vorschlagen. Exemplarisch zeigen auch sie, wie integrative Forschungsinfrastruktursysteme genutzt werden können, um neuartige Fragestellungen effektiv zu bearbeiten.

Alexander **Mehler**, Wahed **Hemati**, Rüdiger **Gleim** und Frank **Baumartz** stellen die Entwicklung von Forschungsinfrastrukturen in den Kontext genereller Digitalisierungstendenzen und zeigen, wie man dies als einen evolutionären Prozess zu neuartigen Systemen auffassen kann. Neben Infrastrukturen zur Visualisierung von Korpusanalyseergebnissen betrachten sie Infrastruktursysteme für linguistische Netzwerke, die in Gestalt von Wikipedia neue Möglichkeiten der Netzwerkanalyse sprachlicher Kommunikation eröffnen.

Im letzten Beitrag dieses Teils wenden sich Hans-Jürgen **Bucher** und Philipp **Niemann** der Medienwissenschaft zu, in der zwar gesprochene oder schriftliche sprachliche Daten eine wichtige Rolle spielen, dies aber eingebettet in eine Vielzahl anderer Modalitäten und Medien. Sie weisen auf einen Nachholbedarf von Infrastrukturen für die Medienforschung hin und zeigen am Beispiel der qualitativen Rezeptionsanalyse, wie durch kleine Forschungseinheiten und einen realistischen Umgang mit Standardisierungserwartungen in Verbindung mit einem Stufenmodell der Entwicklung von Infrastrukturen Forschungsmöglichkeiten geschaffen werden können, die auch bei solchen Erkenntnisinteressen einen erheblichen Mehrwert für die Forschung versprechen.

2.3 Kapitel III – Korpora und Informationssysteme

Im dritten Teil dieses Bandes werden einige ganz bestimmte Korpora und Informationssysteme mit ihren Eigenschaften und in ihrer Genese betrachtet. Den Auftakt dazu machen Ruxandra **Cosma** und Marc **Kupietz** mit einer Darstel-

lung von Korpora und der Korpusinfrastruktur am Institut für Deutsche Sprache, bei der sie eine Parallele zum Infrastrukturbereich des Schienenverkehrs ziehen. Mit der Digitalisierung wird das „Gleissystem“ ausgebaut und die „Geschwindigkeit“ der „Züge“ größer, so dass leistungsfähige Netze entstehen, an denen das IDS maßgeblich beteiligt ist. Aus dem deutschen Referenzkorpus erwächst inzwischen der Plan eines parallelen europäischen Referenzkorpus, dessen Entwicklung mit dem Sprachpaar Deutsch-Rumänisch bereits begonnen worden ist.

Ein zweites Korpus, das von einer kompletten Korpusinfrastruktur umgeben ist, stellen Alexander **Geyken**, Matthias **Boenig**, Susanne **Haaf**, Bryan **Jurish**, Christian **Thomas** und Frank **Wiegand** vor. Für das Deutsche Text-Archiv (DTA) wurden verschiedene Werkzeuge zur Erstellung und Annotation von Textressourcen entwickelt, die durch eine Umgebung zur kollaborativen Qualitätssicherung ergänzt werden. Auch für die Datenanalyse wurden DTA-spezifische Visualisierungsmöglichkeiten für historische Wortverläufe und Kollokationen geschaffen. Da diese Arbeiten parallel zum Aufbau des CLARIN-Verbundes stattgefunden haben und mit diesem abgestimmt wurden, können nach offizieller Beendigung des Projekts sämtliche Angebote im Rahmen von CLARIN weitergeführt werden.

Andrea **Rapp** verlängert in ihrem Beitrag die historischen Linien bis ins Mittelalter. Sie erläutert die integrative Kraft, die die kollaborative Arbeit an Quellensammlungen, Korpora und Wörterbüchern für die Mediävistik aufweist. Die digitale Bearbeitung historischer Quellen gliedert sich dabei in eine Traditionslinie ein, die zur Ausprägung des Forschungsgebietes der *Digital Humanities* geführt hat.

Martine **Dalmas** und Roman **Schneider** befassen sich das Kapitel abschließend mit einem anderen Typ digitaler Sprachressourcen, mit Online-Grammatiken. Das weit ausgebaute Angebot des IDS bietet für sie die Grundlage für die Erörterung der Frage, wie digitale grammatische Informationssysteme insbesondere aus Sicht der Auslandsgermanistik eingesetzt werden können und welche Erwartungen dabei bestehen. Sie betonen, dass grammatische Traditionen in der Kontrastsprache einerseits, strukturelle Differenzen zwischen den Sprachen andererseits dazu führen müssen, die spezifische Perspektive von Forschenden und Sprachlernenden mit einem anderen erstsprachlichen Hintergrund zu berücksichtigen.

2.4 Kapitel IV – Annotation und Modellierung

Im letzten Kapitel des vorliegenden Bandes wird der Bogen beendet, der mit dem ersten Kapitel begonnen wurde. Um funktionierende Kooperationen und

Verbünde zu ermöglichen, ist es notwendig, Daten in standardisierter Form mit Zusatzinformationen anzureichern und die entstehenden Datenstrukturen durch Regeln zu beschreiben, so dass die aufwändig entwickelten Verarbeitungsverfahren auch auf zukünftige Daten angewandt werden können. Diesen Aspekt von Annotation und Modellierung führt C. M. **Sperberg-McQueen** an Hand der *Extensible Markup Language* (XML) aus. In XML lassen sich alle Elemente für die Gewährleistung von Interoperabilität finden: eine definierte Syntax der Annotation, ein definiertes Datenmodell und die Möglichkeit, mit einer „Datengrammatik" die Korrektheit der Annotation zu überprüfen. Anforderungen an die Interoperabilität bestehen aber auch in einem weitergehenden inhaltlichen Sinne hinsichtlich der Datenstrukturierung.

Michael **Beißwenger** zeigt in seinem Beitrag, wie die für textbezogene Forschung in den Geisteswissenschaften entwickelten Dokumentgrammatiken der *Text Encoding Initiative* für Kommunikate der internetbasierten Kommunikation erweitert werden können. Dieser Kommunikationstyp eröffnet aufgrund seiner Unmittelbarkeit und der prinzipiellen Vollständigkeit seiner Erfassung eine interessante Forschungsperspektive für die germanistische Linguistik, führt aber auch zu praktischen Erfassungs- und Annotationsproblemen, zu deren Behebung die auf traditionellen Texttypen entwickelten Verfahren angepasst werden müssen.

Abschließend vollziehen Gerhard **Heyer**, Gregor **Wiedemann** und Andreas **Niekler** den Übergang in die Semantik-Modellierung: Sie zeigen in ihrem Beitrag, wie mit dem Konzept des *Topic Modeling* mit statistischen Mitteln Themen in Texten identifiziert und in ihrer Entwicklung in einem Korpus verfolgt werden können. Der Aspekt der Modellierung tritt dabei in einem erweiterten Sinne in Erscheinung: Nicht nur die Strukturen der Annotation sind Gegenstand der Modellierung und werden als solche auf den Text übertragen, vielmehr werden aus dem Text selbst Strukturen extrahiert, die als Grundlage für weitergehende Analysen fungieren.

3 Perspektiven

Die Vision einer vollständigen Interoperabilität sämtlicher Forschungsdaten mit all ihren Metadaten ist noch lange nicht erreicht. In den letzten Jahren wurden jedoch viele wichtige Fortschritte erzielt, wie die Beiträge in diesem Band zeigen. Als wichtigste Aufgabe für die Zukunft wird es sich erweisen, die entstanden Infrastrukturverbünde langfristig in ihrer Existenz abzusichern und dadurch eine konzertierte Weiterentwicklung der Technologien zu gewährleisten. Auch erweiterte Möglichkeiten der kooperativen Arbeit an den Ressourcen

selbst sowie an den empirischen und qualitativen Ergebnissen ihrer Nutzung stellen ein wesentliches Desiderat dar. Für all das, was in der Vergangenheit bereits geleistet worden ist und was zukünftig noch geleistet werden muss, hat Ludwig M. Eichinger mit seiner Tätigkeit am Institut für Deutsche Sprache in Mannheim wesentliche Grundlagen gelegt.

I Kooperation und Verbünde

Thomas Gloning

1 Forschungsinfrastrukturen und Informationssysteme im Zeichen der Digitalisierung: Aspekte der Kollaboration und der Nutzer-Einbindung

Abstract: Mit der Digitalisierung sind weitreichende Veränderungen der germanistischen Forschung und Lehre verbunden. In diesem Beitrag skizziere ich zunächst wesentliche Veränderungen, die sich aus der Verfügbarkeit digitaler Daten und Werkzeuge ergeben haben. Ich behandle dann die Frage, wie Infrastrukturverbünde zur Gestaltung der Arbeitslandschaft und der Forschungsmöglichkeiten beitragen können und wie fachliche Nutzergruppen hierbei produktiv eingebunden werden können.

Keywords: Germanistik, Digitalisierung, Forschungsinfrastrukturen, Nutzereinbindung, CLARIN-D, DARIAH-DE

1 Einleitung

In den letzten beiden Jahrzehnten haben sich die Arbeitsbedingungen in vielen Bereichen der germanistischen Forschung und Lehre dramatisch verändert. Im Zeichen der Digitalisierung sind nun viele Arten und große Mengen von Sprach-Daten auch elektronisch verfügbar, z. B. historische Texte, Ton- und Videoaufnahmen, gegenwartssprachliche Korpora zu gesprochener, geschriebener und computervermittelter Sprache. Gleichzeitig sind neue und mächtige Werkzeuge für die Erzeugung und die Analyse digitaler Daten verfügbar. Im Bereich der wissenschaftlichen Kommunikation, Kollaboration und Präsentation dienen Werkzeuge wie E-Mail, Mailinglisten, viele Arten von Social Media und Videokonferenztools, aber auch Präsentationsmittel wie PowerPoint und seine Alternativen sowie eine Vielzahl weiterer Angebote der Bewältigung un-

Anmerkung: Für ihre Unterstützung danke ich sehr herzlich den Herausgebern, Henning Lobin, Roman Schneider und Andreas Witt, sowie Melanie Grumt Suárez (CLARIN-D).

Thomas Gloning, Institut für Germanistik, Justus-Liebig-Universität Gießen, D-35394 Gießen, E-Mail: thomas.gloning@uni-giessen.de

https://doi.org/10.1515/9783110538663-002

terschiedlichster wissenschaftlicher Aufgaben. In der akademischen Lehre sind digitale Plattformen eine wesentliche Grundlage für die Organisation und die Zusammenarbeit. Auch die Publikation und Rezeption wissenschaftlicher Literatur wird zunehmend durch digitale Angebote ergänzt.

Mit der Digitalisierung sind neue Formen der Zusammenarbeit und des „community building“ verbunden, die zum Teil auf den genannten Kommunikationswerkzeugen beruhen, aber zum Teil auch durch gemeinsame Interessen bedingt sind. Thematisch fokussierte wissenschaftliche Mailinglisten sind Beispiele für inzwischen etablierte Formen der gemeinschaftlichen Zusammenarbeit und des wissenschaftlichen Austauschs zu bestimmten Themen und Themenfeldern. Aber auch die Abonnements von wissenschaftlichen Blogs oder themenspezifischen Twitter-Accounts sind neuartige Formen der Gemeinschaftsbildung, die teilweise eigene Spielarten der Vernetzung hervorbringen können. Unterschiedliche Formen des Community Building wurden und werden darüber hinaus im Rahmen des Aufbaus von Infrastrukturverbünden planvoll betrieben. So haben zum Beispiel die großen Infrastrukturprojekte CLARIN-D und DARIAH-DE eigene Formate der Einbindung von Nutzergruppen entwickelt und implementiert.[1]

Während viele Germanistinnen und Germanisten daneben nach wie vor und je nach Forschungsgegenstand und Zielsetzung mit gutem Recht traditionelle Methoden der germanistischen Forschung anwenden und keine digitalen Ressourcen einsetzen, gehören digitale Daten und Werkzeuge doch zunehmend zum wissenschaftlichen Alltag. Dabei zeigt sich, dass digitale Sprachdaten und Werkzeuge für verschiedene Forschungsfragen in unterschiedlicher Weise relevant sind. Wer eine literaturwissenschaftliche Forschungsarbeit zu Rilkes fünfter Duineser Elegie im Kontext der bisherigen Rilke-Forschung oder eine wissenschaftsgeschichtliche Arbeit über Jacob Grimms Stellung zum Fremdwort schreibt, wird eher traditionelle Verfahren nutzen müssen: Texte lesen, Notizen machen, Gliederungsentwürfe schreiben usw. Aber selbst bei einer solchen Arbeit werden inzwischen E-Mail, computergestützte Textverarbeitung und die Nutzung digitaler Bibliotheksangebote eine wichtige Rolle spielen, auch wenn die eigentliche intellektuelle Arbeit vielleicht ohne nennenswerte digitale Anteile stattfindet.

Wer demgegenüber zum Beispiel eine sprachwissenschaftlich-diskursorientierte Forschungsarbeit zum Wortgebrauch im Flüchtlings- und Migrationsdis-

---

1 In CLARIN-D sind dies u. a. die Facharbeitsgruppen: https://www.clarin-d.net/de/facharbeitsgruppen (20. 1. 2018). In DARIAH gehören sog. „Working Groups“ dazu: https://www.dariah.eu/activities/working-groups-list (20. 1. 2018). Aber auch Instrumente wie Befragungen, Workshops, Schulungen usw. werden hierfür genutzt.

kurs der Jahre 2015–2017 schreibt, der wird um die Nutzung der elektronischen Korpora des Instituts für deutsche Sprache,[2] der Berlin-Brandenburgischen Akademie der Wissenschaften[3] und ggf. auch des Leipziger Wortschatzportals[4] nicht herumkommen. Und wer eine literaturwissenschaftliche Arbeit zu Rilke im Kontext seiner Zeit in einem „Digital Humanities"-Kontext schreiben will, wird digitale Textressourcen und entsprechende Tools in sachgemäßer Weise verbinden müssen. Im Bereich der Literaturwissenschaften gibt es inzwischen eine avancierte digital unterstützte Forschung, exemplarisch sind etwa die Arbeiten zu nennen, über die in loser Folge in den „pamphlets" des Stanford Literary Lab berichtet wird.[5] In der germanistischen Literaturwissenschaft sind, um nur drei Namen zu nennen, etwa die Arbeiten von Fotis Jannidis, Gerhard Lauer oder Jan-Christoph Meister zu erwähnen. Diese und andere Literaturwissenschaftlerinnen und Literaturwissenschaftler sind mit zahlreichen Forscherinnen und Forschern aus anderen (Teil-)Disziplinen seit 2012 im Verband „Digital Humanities im deutschsprachigen Raum" (DHd) zusammengeschlossen. Jahrestagungen, zahlreiche Projekte und eine Internetseite mit Blog dokumentieren die Konsolidierung dieser Fachgemeinschaft.[6]

Viele neue Forschungsprojekte, die zum Teil durch das BMBF, die DFG oder durch Stiftungen in strukturierten Programmen gefördert werden, setzen sich zum Ziel, neuartige Nutzungen von digitalen Daten und Werkzeugen für traditionelle oder neuartige Fragestellungen zu entwickeln, zu erproben und zu evaluieren. Und auch im Rahmen dieser Programme ist die Vernetzung unter den Projekten eine wichtige Komponente.

Vor dem Hintergrund dieser weitreichenden Entwicklungen behandle ich nachfolgend – aus einer durchaus persönlichen Perspektive – folgende Fragestellungen und Teilthemen:

– Wie haben sich Kernbereiche der germanistischen Arbeitslandschaft durch die Digitalisierung verändert? Welche (alten und neuen) Formen der Zusammenarbeit, der „Vergemeinschaftung" und des „Community Building" sind zu beobachten? (Abschnitt 2)
– Wie kann man die Landschaft der digitalen Forschungsinfrastrukturen und Informationssysteme mit Bezug zur Germanistik, wie sie sich in den letzten Jahren und Jahrzehnten entwickelt hat, in ihren Grundzügen charakteri-

2 http://www.ids-mannheim.de/kl/projekte/korpora.html (20. 1. 2018).
3 https://www.dwds.de/d/korpora (20. 1. 2018).
4 http://wortschatz.uni-leipzig.de/ (20. 1. 2018).
5 https://litlab.stanford.edu/; https://litlab.stanford.edu/pamphlets/ (11. 11. 2017); vgl. Moretti (2007, 2013).
6 https://dig-hum.de/ (11. 11. 2017); http://dhd-blog.org (18. 1. 2018).

sieren? Wie tragen die einzelnen Angebote zu einem systematischen Ausbau der Arbeits- und Forschungsmöglichkeiten im Bereich der Germanistik bei? (Abschnitt 3)
- Welche Perspektiven und Zukunftsaufgaben ergeben sich im Bereich der Germanistik, die ja durch eine äußerst vielgestaltige innere Differenzierung gekennzeichnet ist, für den weiteren Ausbau von Infrastrukturverbünden wie CLARIN-D und DARIAH-DE? (Abschnitt 4)

2 Kollaborative Arbeitsformen und Formen der „Vergemeinschaftung" in der Germanistik: Was ändert sich mit der Digitalisierung?

Wissenschaft ist auf koordinierte Zusammenarbeit angewiesen. Abhängig von ihrem übergeordneten Zweck haben sich Strukturen zum einen in evolutionärer Weise herausgebildet, zum anderen waren und sind sie Resultat von gezielten Planungsprozessen. Dabei lassen sich mehrere zentrale Funktionskreise unterscheiden, die man im Hinblick auf Veränderungen im Rahmen der Digitalisierung betrachten kann. Diese Veränderungen weisen freilich unterschiedliche zeitliche Dynamiken in den germanistischen Teildisziplinen auf. Mit der Digitalisierung gehen auch neue Entwicklungen in den Formen der „Vergemeinschaftung" einher. Schließlich kann man fragen, ob und inwiefern es Konvergenzen und Unterschiede mit anderen Disziplinen, z. B. anderen Einzelphilologien oder auch im Rahmen der Sprach-, Literatur- und Kulturwissenschaften gibt.

Wenn man von einer allgemeinen Tendenz wie der Digitalisierung in einer breit differenzierten wissenschaftlichen Disziplin wie der Germanistik spricht, dann kann man diese Tendenz in einem ersten Schritt konkretisieren, indem man die unterschiedlichen Funktionskreise betrachtet, in denen sich Resultate der Digitalisierung zeigen.

Einen ersten Kernbereich stellen Formen des wissenschaftlichen Austauschs dar. Der wissenschaftliche Austausch und die kollegiale fachliche Diskussion fand traditionell unter anderem in persönlichen oder telefonisch geführten Gesprächen, auf Tagungen, über Briefwechsel[7] und den Austausch von Sonder-

7 Viele Gelehrtenbriefwechsel aus dem fachlichen Bereich der Germanistik zeugen davon, wie wichtig Briefe für den wissenschaftlichen Austausch zwischen Personen waren, die oft an unterschiedlichen Universitäten wirkten und die aufgrund der geringen Größe von Instituten nicht immer einen fachlichen Gesprächspartner an der eigenen Universität hatten. – „Ein

drucken statt. Mit der Digitalisierung kamen zunächst neue technisch-mediale Formate wie E-Mail, Mailinglisten, Videokonferenzen, Projektmanagement-Software oder spezifische Werkzeuge für den Datenaustausch hinzu. Über die individuelle Nutzung dieser medialen Angebote hinaus bildeten sich für unterschiedliche Gemeinschaften und Arbeitszusammenhänge evolutionär Kombinationen heraus oder wurden planvoll zu sinnvollen Konstellationen zusammengestellt. Prototypisch sind hierfür etwa die Arbeitsinfrastrukturen kleinerer und mittlerer Forschungsprojekte, zu denen im Hinblick auf den Austausch häufig eine Mailingliste, ein Projektmanagementsystem, ein digitales Repositorium für gemeinsam genutzte Daten, eine Umgebung für kollaborative Textproduktion und eine Videokonferenzumgebung für virtuelle Treffen gehören. Zeitlich nur vorübergehend sind etwa die Formen des konferenzbegleitenden Twitterns, das über eigene Hashtags organisiert wird und eine parallele, nicht an den Raum der Konferenz gebundene Kommunikationswelt eröffnet. Man sieht an diesen Beispielen gut, dass mit der koordinierten Nutzung digitaler Werkzeuge neue Formen der „Vergemeinschaftung" und neue Sozialformen entstanden sind, denen wir uns unten ausführlicher zuwenden.

Ein zweiter Funktionskreis bezieht sich auf wissenschaftliche und organisatorische Informationen, zum Beispiel zu geplanten oder stattgefundenen Tagungen, zu Stellenausschreibungen, zu den gesetzlichen Grundlagen der Wissenschaft, zur Bibliographie des Fachs und seiner Bereiche usw. In der vordigitalen Wissenschaft wurden für diese Zwecke wiederum (Reihen-)Briefe für den Austausch verwendet, aber auch in den Fachzeitschriften fanden sich feste Rubriken zu Informationen dieser Art. Heute dienen zusätzlich Mailinglisten (z. B. www. linguistlist.org), fachliche Portale wie zum Beispiel HSozKult[8] und „Germanistik im Netz",[9] fachbibliographische Webangebote wie die IDS-Seite „Bibliographien zur Linguistik/Literaturlisten"[10] oder eigene Stellenportale für die Wissenschaft solchen Zwecken. Ein Problem an diesen Entwicklungen

---

Fluch unserer Wissenschaft ist die Isolirung der Fachgenossen und dem sollen doch die Congresse abhelfen" schrieb der Islamwissenschaftler Ignaz Goldziher 1894 an seinen Kollegen Hartmann (Hanisch 2000: 13), nachdem er ihn in Briefen immer wieder bestärkt hatte, einen anstehenden Kongress in Genf zu besuchen. Das Beispiel, das sich gewiss nicht ohne Weiteres auf die germanistischen Verhältnisse der Zeit übertragen lässt, zeigt aber doch, dass die Briefwechsel und die gelegentlichen Zusammentreffen auf Kongressen in den vergangenen Jahrzehnten und Jahrhunderten ein wesentliches Mittel des wissenschaftlichen Austauschs waren, dabei aber einem noch unbeschleunigten Zeittakt folgten.

8 http://www.hsozkult.de (H/SOZ/KULT – Kommunikation und Fachinformation für die Geschichtswissenschaften); (8. 10. 2017).

9 http://www.germanistikimnetz.de/neuerscheinungen/ (15. 1. 2018).

10 http://www.ids-mannheim.de/service/quellen/biblio.html (8. 10. 2017).

könnte man darin sehen, dass es inzwischen in vielen Bereichen keineswegs einfach ist, die Übersicht, was für Angebote zur Germanistik und ihren Fachzonen es gibt, zu bekommen oder zu behalten. Auf dieses Problem sind wiederum spezielle Überblicksangebote bezogen, die den Nutzern diese Übersicht verschaffen und aktuell halten sollen. Im Bereich der Altgermanistik könnte man hier zum Beispiel das Mediaevum-Portal[11] nennen, für den Gesamtbereich der Germanistik das von der DFG geförderte Portal „Germanistik im Netz",[12] ein umfangreiches internationales Germanist/-innen-Verzeichnis wird auf der Seite des Deutschen Germanistenverbandes angeboten.[13]

Ein dritter Funktionsbereich ist das Publikationswesen. Die Veröffentlichung von Resultaten der Wissenschaft ist eines ihrer Kerngebote (Weinrich 1995a, b), gleichzeitig führt die Auffächerung in zahlreiche Disziplinen zu einer Vielfalt von spezifischen Publikationskulturen. Die Germanistik ist traditionell geprägt von einem zweigleisigen System, das Monographien und Zeitschriftenaufsätze gleichermaßen umfasst, wobei diese beiden Veröffentlichungsarten unterschiedliche und komplementäre Rollen spielen im Hinblick auf fachliche Ziele und die akademischen Karrierewege ihrer Verfasser/-innen. Lehrbücher, Handbücher, Bibliographien treten hinzu, darüber hinaus gibt es spezifische Publikationsformen, die mit den sprachlich verfassten Gegenständen des Fachs zu tun haben: Historische Wörterbücher und kommentierte Texteditionen sind vielleicht die prominentesten Beispiele, für beide Bereiche gibt es eine hochspezialisierte Methodenlehre und -diskussion sowie ein breites Spektrum digitaler Angebote.[14] Im Hinblick auf Aspekte der Digitalisierung ist das germanistische Publikationswesen zur Stunde gleichzeitig durch Aspekte der Beharrung und der Innovation gekennzeichnet, deren Verhältnis sich nicht in allgemeiner Weise charakterisieren lässt.

Ein eigener Bereich innerhalb des Publikationswesens sind die Rezensionen. Sie sind einerseits *Teil* der Publikationslandschaft, sie sind andererseits reflexiv auf wissenschaftliche Publikationen und Fragen der Qualitätssicherung bezogen. Die beiden Hauptfunktionen von Rezensionen sind das Informieren über neue Publikationen und die kritische Beurteilung des jeweiligen

---

11 http://www.mediaevum.de (8. 10. 2017).

12 http://www.germanistik-im-netz.de (8. 10. 2017).

13 http://www.germanistenverzeichnis.phil.uni-erlangen.de/ (8. 10. 2017).

14 http://www.woerterbuchnetz.de (15. 1. 2018); https://www.dwds.de (20. 1. 2018). Zur digitalen Editorik vgl. exemplarisch Bender 2016; Sahle 2013. Vgl. weiter die Angebote des „Institut für Dokumentologie und Editorik" (https://www.i-d-e.de), das dazugehörige Review Journal RIDE (http://ride.i-d-e.de; beide 20. 1. 2018) sowie das „Magazin für digitale Editionswissenschaften", das vom Interdisziplinären Zentrum für Editionswissenschaften der FAU Erlangen-Nürnberg herausgegeben wird.

Werks im Hinblick auf spezifische fachliche Kriterien. Die traditionelle Textform für diese kommunikative Aufgabe ist die Rezension, die im späten 17. Jahrhundert mit der Entstehung der wissenschaftlichen Zeitschriften aufgekommen ist und dann im 18. Jahrhundert eine erste Blüte erreichte (Habel 2007). Rezensionen sind eine stabile Komponente der Wissenschaftskommunikation auch im Bereich der Germanistik, die sich vor allem in den Fachzeitschriften der Germanistik und ihrer Teildisziplinen entfaltet hat. Mit der Digitalisierung wurden, soweit derzeit erkennbar, mehrheitlich die traditionellen Darstellungsformen im Umkreis der Rezension im neuen Medium weiter verwendet. Neuartig sind Rezensionen, die auf digitale Angebote, z. B. Editionen, bezogen sind und neue Kriterien der Beurteilung erfordern.[15]

Der Funktionswandel im Bereich der Literaturversorgung ist in erster Linie an den Strukturwandel im Bereich der Bibliotheken, der Verlage, teilweise auch der Gesetzgebung, den Strategien des Erwerbs von Konsortial-Lizenzen (z. B. DFG-Nationallizenzen) und nicht zuletzt auch Open-Access-Prinzipien gebunden. Die meisten Universitätsbibliotheken, aber auch das Institut für Deutsche Sprache (IDS) und das „Germanistik im Netz“-Portal betreiben heute digitale Publikationsserver und stellen ein breites digitales Literaturangebot für die Nutzer/-innen bereit. Bibliotheksnahe Digitalisierungszentren, zum Beispiel in München oder Göttingen, setzen strategische und mit öffentlichen Mitteln geförderte Programme um, so etwa die Digitalisierung der alten Drucke aus dem 16., 17. und 18. Jahrhundert, wie sie in den VD16-, VD17- und VD18-Verzeichnissen dokumentiert sind.[16] Bei der Retrodigitalisierung und der Vernetzung der Wörterbücher des Deutschen sowie bei vielen digitalen Editionen hat das „Trier Center for Digital Humanities“ wesentliche Beiträge geleistet.[17] Die Tatsache, dass heute alte Drucke und Handschriften digital verfügbar und durch die Erschließung von Metadaten auch systematisch suchbar und auffindbar sind (etwa über die KVK-Suchmaschine),[18] stellt einen Quantensprung in den Arbeitsbedingungen der sprach- und literarhistorischen Forschung dar, den vielleicht nur die Personen angemessen beurteilen und würdigen können, die noch mit dem zähen Verfahren der Bestellung eines Mikrofilms per Brief,

15 Vgl. z. B. die in der vorhergehenden Fußnote genannten Zeitschriften.

16 www.vd16.de; www.vd17.de; www.vd18.de (11. 11. 2017); von diesen Kurzadressen aus wird man jeweils weitergeleitet auf die entsprechenden Katalogstartseiten.

17 http://kompetenzzentrum.uni-trier.de (20. 1. 2018); http://woerterbuchnetz.de (20. 1. 2018).

18 https://kvk.bibliothek.kit.edu (11. 11. 2017) mit der Option „Nur digitale Medien suchen“. Digitalisate von Handschriften sind u. a. bei www.manuscripta-mediaevalia.de (15. 1. 2018) und im Handschriftencensus (www.handschriftencensus.de; 15. 1. 2018) verzeichnet. Eine systematische und breite Dokumentation von Angeboten zu historischen Sprachdaten und zu ihren Erschließungsmitteln kann an dieser Stelle nicht geleistet werden.

der achtwöchigen Wartezeit, der freudig begrüßten Ankunft der Mikrofilm-Sendung, dem Ausfüllen der Überweisung für den Repro-Auftrag, den langweiligen Stunden vor dem Mikrofilm-Rückvergrößerungsgerät usw. vertraut sind. Im Hinblick auf historische Daten ist es besonders wichtig, dass viele Texte heute nicht nur als Bilddigitalisate, sondern auch als standardisiert erfasste, interoperabel nutzbare, maschinell durchsuchbare und bearbeitbare Volltexte verfügbar sind.

Eine weitere wichtige Entwicklung ist die Ablösung der älteren „Sondersammelgebiete“ durch fachlich spezialisierte Forschungsinformationsdienste, mit denen die beauftragen Bibliotheken nicht nur forschungsbezogene (gedruckte und elektronisch verfügbare) Literatur, sondern auch digitale Forschungsdaten verzeichnen und teilweise selbst anbieten. Die Dokumentation der Angebote erfolgt digital und dynamisch anpassbar in fachlichen Portalen, bei der technischen Umsetzung werden avancierte Metadaten-, Linked-Open-Data-, Abfrage- und andere Technologien verwendet. Für die germanistische Forschung wird der Ende 2017 bewilligte „Fachinformationsdienst Germanistik“ einschlägig sein, der auch die Arbeiten am Portal „Germanistik im Netz“ weiterführt, die germanistische Sprachwissenschaft wird sicherlich auch vom „Fachinformationsdienst Linguistik“ und dem Portal „linguistik.de“ profitieren, dasselbe gilt für die germanistische Literaturwissenschaft und den „Fachinformationsdienst Allgemeine und Vergleichende Literaturwissenschaft“.[19]

Von all diesen Entwicklungen ist auch die Hochschul-Lehre im inneren Kern betroffen: Die Digitalisierung der Materialien, der Präsentationsweisen, der Lektüretexte, der Textproduktion bis hin zur digitalen Verbreitung von Vorlesungen usw. prägt heute viele Bereiche in der Lehre an den Universitäten. Hochschuldidaktische Planungen und Konzeptionen müssen folglich die Frage nach dem (Nicht-)Einsatz digitaler Ressourcen mit einbeziehen. Digitale Lernplattformen wie StudIP, Ilias oder Moodle sind häufig eingesetzte Umgebungen für die Nutzung digitaler Angebote und digital gestützter Verfahrensweisen der Zusammenarbeit und der Präsentation. In einem sprach- und literaturbezogenen Fach wie der Germanistik sind digitale Forschungsmethoden seit vielen Jahren auch Gegenstand der Lehre, z. B. im Bereich Korpuslinguistik.

Freilich sind derzeit noch deutliche Abstufungen im Hinblick auf Art, Umfang und Funktion des Einsatzes digitaler Daten, Werkzeuge und Lehrkomponenten im weitesten Sinne zu erkennen. Sie hängen zum Teil mit den Lehrgegenständen zusammen, teilweise aber auch mit persönlichen Einstel-

19 FID Germanistik: https://www.ub.uni-frankfurt.de/ssg/dsl.html; FID Allgemeine und Vergleichende Literaturwissenschaft: https://www.ub.uni-frankfurt.de/projekte/avl.html; FID Linguistik: https://www.ub.uni-frankfurt.de/projekte/fid-linguistik.html (alle 15. 1. 2018).

lungen von Lehrenden. In den verschiedenen germanistischen Teildisziplinen sind unterschiedliche zeitliche Dynamiken von Aspekten der Digitalisierung zu beobachten, auch wenn sie insgesamt schwer systematisierbar und empirisch derzeit nicht zu belegen sind. So gibt es Fachkulturen, in denen Tagungsvorträge eher vorgelesene Manuskripte, andere, in denen sie üblicherweise frei gesprochene Präsentations-Aufführungen sind.[20] Im Hinblick auf die Präferenzen von Papier- vs. PDF-Lesetexten scheint es bei Forscher/-innen, Lehrenden und Studierenden unterschiedliche Abschattungen zu geben.[21]

Die Digitalisierung hat auch neue Formen bzw. neue Realisierungsweisen der „Vergemeinschaftung“ in der Germanistik hervorgebracht. Traditionelle Formen sind zum Beispiel persönliche Netzwerke, die auf institutioneller Zugehörigkeit, gemeinsamen Interessen oder geteilten Aufgaben beruhen.[22] Temporäre Formen sind zum Beispiel Forschungsprojekte, die nur für den Zeitraum ihrer Durchführung Bestand haben. Im Hinblick auf Fachverbände wie den „Germanistenverband“, die „Deutsche Gesellschaft für Sprachwissenschaft“ oder die „Gesellschaft für germanistische Sprachgeschichte“ kann man eine institutionelle Perspektive einnehmen und Fragen stellen wie die nach den Zielen, der Dauer des Bestehens, der Entwicklung von Mitgliederzahlen oder nach dem Profil der Aktivitäten. Aus einer individuellen Perspektive stellt sich z. B. die Frage, welche Rolle die Mitgliedschaft für einzelne Wissenschaftler/-innen spielt, sei es in wissenschaftlicher Hinsicht, sei es für die berufliche Karriereentwicklung.

Mit der Digitalisierung sind diese traditionellen Formen der Vergemeinschaftung nicht weggefallen, auch Realisierungsformen wie Jahrestagungen von Gesellschaften erfreuen sich weiterhin großer Beliebtheit. Aber es sind neue Formen auf digitaler Grundlage hinzugekommen, z. B. wenn Fachgesellschaften ihre Mitglieder über eine Mailingliste informieren oder einen regelmäßigen digital verschickten Newsletter haben. Auch Fachgesellschaften präsen-

---

20 Vgl. zu Präsentationen und ihren Vorläufern Lobin (2009), Schnettler & Knoblauch (2007).

21 Auf der Humanist-List berichteten mehrere Mitglieder auf die Frage nach „sustained reading from screen“, also nach dauerhafter Lektüre am Bildschirm, dass sie ihre Lektüre- und Annotationsumgebung komplett auf digitale Werkzeuge umgestellt haben. Umgekehrt bekunden nicht wenige der Studierenden, dass sie lieber „richtige“ Bücher lesen und digitale Textauszüge im Hinblick auf die Zugänglichkeit zwar schätzen, sie dann aber ausdrucken um sie zu lesen und zu bearbeiten. Die Ausgangsmail zur Umfrage „sustained reading from screen“ auf der Humanist-List ist hier archiviert: http://lists.digitalhumanities.org/pipermail/humanist/2017-October/015131.html (11. 11. 2017). Die Ausgangsseite des Listenarchivs: http://lists.digitalhumanities.org/pipermail/humanist/ (11. 11. 2017).

22 In medialer Hinsicht waren die Gelehrtenbriefwechsel für viele Jahrhunderte ein zentrales Werkzeug für den wissenschaftlichen Austausch und die themenzentrierte Vergemeinschaftung.

tieren sich über soziale Medien nach außen, sei es über Internetauftritte, Facebook-Seiten, Twitter oder eigene Blogs. Darüber hinaus sind eigene Fachverbände hinzugekommen, die sich auch Fragen der Digitalisierung widmen, z. B. die Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL).[23]

Sodann sind neue Formen der Vergemeinschaftung entstanden, die an digitale Medien gebunden sind. Mailinglisten sind eine der frühen wissenschaftlichen Nutzungsformen der Email-Technologie, die auch heute noch vielfach genutzt werden und eigene Nutzungsprofile vor allem in den Bereichen Service/Information, Kollaboration und Kritik/Kontroverse aufweisen.[24] Digitale Datenaustausch-Gemeinschaften wie das von Jost Gippert begründete TITUS-Portal (Thesaurus Indogermanischer Text- und Sprachmaterialien) gehören zu den frühen Nutzungen digitaler Technologien, die kontinuierlich weiterentwickelt und angepasst wurden. Man könnte hier noch viele weitere Entwicklungen anführen, die zeigen, dass digitale Werkzeuge, Daten und Verfahrensweisen den Alltag von Wissenschaftler/-innen tiefgreifend verändert haben, dass sie aber im Hinblick auf die epistemischen Kernaufgaben in den verschiedenen Teildisziplinen sehr unterschiedliche Rollen spielen. Diese Veränderungen sind immer wieder auch mit reflexiven Diskussionen verbunden, die sich auf Themen wie die Fortschrittsrhetorik des Digitalen, die Verträglichkeit hermeneutisch-geisteswissenschaftlicher Fragestellungen mit digitalen Zugriffen oder auch die Frage nach mediengeschichtlichen Entwicklungen („Untergang des Buchs") bezogen.[25]

Wir wenden uns nun der Rolle von digitalen Forschungsinfrastrukturen zu, die als ein Bereich strategischer Wissenschaftsplanung im europäischen Zusammenhang gesehen werden müssen.

3 Infrastrukturangebote und Nutzer-Orientierung

Die Veränderungen, die im zweiten Abschnitt skizziert wurden, sind zum einen auf mehr oder weniger individuelle Formen der Nutzung digitaler Angebote zurückzuführen. Ein Wissenschaftler, der einen fachlichen Blog betreibt, eine

23 http://www.gscl.org (20. 1. 2018).

24 Vgl. hierzu Bader (2018), Gloning & Fritz (2011). – Einen interessanten quasi-autobiographischen Beitrag zu Fragen der Akzeptanz digitaler Medien in den frühen 1990er Jahren bietet Krappmann 1993.

25 Vgl. exemplarisch Lobin (2014), Flanders (2009), Deegan & McCarty (2012). McCarty schreibt über die Frühgeschichte des humanities computing: „Historically, the computer came to the humanities from outside and was received as a foreigner, in ‚fear and trembling' (Nold 1975), as well as with curiosity." (McCarty 2012: 8).

Wissenschaftlerin, die eine Fachzone mit einem Twitter-Account betreut, eine wissenschaftliche Institution, die sich ein Facebook-Profil zulegt, dies sind zunächst individuelle Angebote, die so oder anders sein könnten und dem Markt der Medienangebote opportunistisch folgen (oder auch nicht). Zum anderen werden Veränderungen aber auch eröffnet durch digitale Infrastrukturangebote, die das Ergebnis von wissenschaftsstrategischen Planungen auf unterschiedlichen Ebenen mit verschiedenen Reichweiten sein können. Man kann folgende Ebenen und Reichweiten unterscheiden:

– eine lokale Ebene (z. B. eine digitale Lernplattform einer einzelnen Universität);
– eine überregionale Ebene (z. B. ein Konsortialvertrag auf Länderebene für Lizenzen digitaler Ressourcen);
– die nationale Ebene (z. B. DFG-Lizenzen, die nationalen Ableger der europäischen Infrastrukturverbünde);
– die europäische Ebene (z. B. die Infrastrukturverbünde CLARIN und DARIAH);
– die internationale Ebene (z. B. die Text Encoding Initiative[26] als Grundlage für die interoperable, nachhaltige Nutzung digitaler Text-Angebote aus – fast – allen Sprachen und Zeitstufen der Überlieferung oder die Kataloge der European Language Resources Association[27]).

Diese Ebenen-Architektur trägt an manchen Stellen freilich eher zur Verdunkelung wichtiger Zusammenhänge bei, denn schon am Beispiel eines lokalen Publikationsservers etwa einer Universitätsbibliothek sieht man, dass die Initiative und der Betrieb zwar lokal, die Nutzungsmöglichkeiten über die Vermittlung der nationalen Kataloge aber international sind. Auch die bibliographischen Ressourcen des IDS und seine digitalen Publikationsangebote sind auf diese Weise überregional und international auffindbar.

Die stärker „individualisierten“ Entwicklungen im Bereich der Digitalisierung der Wissenschaften sind insgesamt nicht leicht überschaubar. Um sie zu beantworten, müsste man zum Beispiel wissen, welche Rolle ein ganz alltägliches Instrument wie die Google-Suchmaschine in den Arbeitslandschaften von Wissenschaftler/-innen in vielen Disziplinen spielt. Man weiß, dass Suchmaschinen auch aus dem wissenschaftlichen Alltag „nicht mehr wegzudenken“ sind. Aber welche Rolle genau spielen Suchmaschinen in den verschiedenen Fachzonen der Germanistik, wie und wofür werden sie genutzt? Spielt z. B. das Ergebnis-Ranking von Suchmaschinen eine Rolle für die Auswahl von

---

26 TEI: http://www.tei-c.org (20. 1. 2018).
27 ELRA: http://www.elra.info/en/catalogues (20. 1. 2018).

Lehrmaterialien, Lektüretexten (unter dem Diktat knapper Zeit)? Wann wird für sprachbezogene Fragen Google genutzt, wann die wissenschaftlichen Korpora? Vergleichbare Fragen kann man in Bezug auf viele andere digitale Werkzeuge und Angebote in der germanistischen Fachzone stellen.

Im Hinblick auf die Organisation und den Betrieb der Wissenschaft als eine öffentliche Aufgabe ist demgegenüber vor allem die Frage interessant, welchen Beitrag die strategisch geplanten und mit öffentlichen Mitteln finanzierten Maßnahmen zur Verbesserung der digitalen Arbeitslandschaft leisten. Eine erste Leitfrage, mit der man diese Perspektive konkretisieren kann, zielt zunächst darauf, wie man die Landschaft der strategisch geplanten digitalen Forschungsinfrastrukturen und Informationssysteme mit Bezug zur Germanistik, wie sie sich in den letzten Jahren und Jahrzehnten entwickelt haben, in ihren Grundzügen charakterisieren kann. Ich kann hier keinen irgendwie gearteten vollständigen Überblick geben, möchte deshalb nur wenige *Typen* von Angeboten charakterisieren.

Man kann dabei von den Institutionen her denken und zum Beispiel fragen, welchen Beitrag Bibliotheken aktuell leisten und zukünftig leisten können.[28] Digitale Angebote von Bibliotheken umfassen unter anderem mehrdimensional erschlossene und durchsuchbare Kataloge, es gibt darüber hinaus Meta-Kataloge wie den Karlsruher Verbundkatalog, viele Bibliotheken betreiben digitale Publikationsrepositorien und Publikationsserver mit neuen, vielfach im Open-Access-Modus veröffentlichten Werken, die systematische Retrodigitalisierung älterer Werke ist, wie oben erwähnt, vielfach ebenfalls an die Bestände und Leistungen einzelner Bibliotheken gebunden, Fachinformationsdienste als Nachfolge-Einrichtungen zu den älteren Sondersammelgebieten werden von Bibliotheken betreut. Dies ist sicherlich nur ein Ausschnitt aus dem Funktionswandel der Bibliotheken im Zeichen der Digitalisierung. Klar ist aber, dass Bibliotheken zu den Hauptakteuren bei der Gestaltung der weiteren Entwicklung nicht nur der digitalen Literaturversorgung und -erschließung gehören werden, sondern ggf. auch bei der Erschließung und Betreuung von digitalen Forschungsdaten.

Man kann aber auch von den Arten von Angeboten her denken und zum Beispiel fragen: Wie kommen und kamen einzelne fachliche Portale in die Welt? Neben den individuellen Initiativen (z. B. das Bibliotheca Augustana-Portal von Ulrich Harsch) gibt es Angebote, die an einzelne Universitäten gebunden sind oder waren, z. B. das Portal handschriftencensus.de, das zunächst ein Marburger Projekt war und inzwischen in die Akademieförderung auf-

---

28 Der Funktionswandel der Bibliotheken im digitalen Zeitalter wird von einer breiten reflexiven Literatur begleitet, exemplarisch etwa Brown (2016).

genommen wurde, oder das an der Universität Duisburg-Essen angesiedelte LINSE-Portal.[29] Sodann gibt es aber auch die strategisch geplanten und geförderten Portale wie z. B. „linguistik.de“ oder „germanistikimnetz.de“, die Bestandteil der strategischen Entwicklung der Literaturversorgung und der Ressourcendokumentation im Rahmen des FID-Programms der Deutschen Forschungsgemeinschaft sind. Als Beispiele für stärker thematisch fokussierte Angebote kann man das grammatische Informationssystem GRAMMIS[30] oder das „Informationsportal Gesprächsforschung“[31] nennen.

Eine besonders wichtige Rolle für den Ausbau der Versorgung mit digitalen Daten, Werkzeugen und der Erschließung durch Metadaten spielen die Infrastrukturverbünde, die im Rahmen von BMBF-Projektförderungen aufgebaut wurden. Die beiden großen Infrastrukturverbünde CLARIN-D[32] und DARIAH-DE[33] sind Teil der europäischen Wissenschaftsplanung, die im Rahmen des übergeordneten Horizon2020-Programms[34] in sog. ESFRI-Roadmaps festgehalten sind.[35] Ziel ist es, national geförderte Forschungs-Infrastrukturen aufzubauen, die in einem europäischen Kontext zusammenarbeiten. Für die europäischen ESFRI-Projekte wie CLARIN-EU und DARIAH-EU wurde eine eigene konsortiale Rechtsform geschaffen, das „European Research Infrastructure Consortium“ (ERIC). Die beiden Initiativen sind inzwischen in fast allen europäischen Ländern vertreten, in den Niederlanden haben CLARIN und DARIAH fusioniert.[36] Die DARIAH-Initiative ist stärker orientiert an Projekten, die Ressourcen entwickeln, im CLARIN-Kontext spielen verbundene Zentren, die digitale Daten und Werkzeuge bereitstellen, die wesentliche Rolle.[37] Die Angebote von DARIAH-DE und CLARIN-D sind in eigenen Beiträgen dieses Bandes besprochen.

---

29 Linguistik-Server Essen: http://www.linse.uni-due.de (17. 1. 2018).

30 http://grammis.ids-mannheim.de (20. 1. 2018); siehe dazu den Beitrag von Dalmas & Schneider in diesem Band sowie Schneider & Schwinn (2014).

31 http://gespraechsforschung.de (20. 1. 2018).

32 https://www.clarin-d.net/de/ (12. 11. 2017).

33 https://de.dariah.eu/ (12. 11. 2017).

34 http://ec.europa.eu/programmes/horizon2020/ (12. 11. 2017).

35 http://www.esfri.eu (European Strategy Forum on Research Infrastructures); für die ersten Roadmaps: http://www.esfri.eu/roadmap-archive (12. 11. 2017).

36 https://www.clariah.nl/ (12. 11. 2017).

37 Die CLARIN-D-Zentren: https://www.clarin-d.net/de/ueber/zentren. – Beispiele für Werkzeuge: Die „Federated Content Search“ erlaubt die Suche nach Ausdrücken im vernetzten Angebot der Zentren: https://www.clarin-d.net/de/auffinden/fcs-suche-in-ressourcen (20. 1. 2018). – WebLicht ist eine serviceorientierte Umgebung für die Erstellung annotierter Textkorpora, in der unterschiedliche Werkzeuge baukastenartig kombiniert werden können: https://www.clarin-d.de/de/sprachressourcen-und-dienste/weblicht (20. 1. 2018).

Im Hinblick auf die Nutzer/-innen und die vielfältigen wissenschaftlichen Aufgaben, die sie verfolgen, lässt sich eine zweite Leitfrage formulieren: Wie tragen die einzelnen digitalen Angebote zu einem systematischen Ausbau der Arbeits- und Forschungsmöglichkeiten im Bereich der Germanistik bei? Für die Beantwortung dieser Frage können wir uns, wiederum exemplarisch, an einzelnen zentralen Funktionskreisen wissenschaftlicher Arbeit orientieren. Ich will dies an drei germanistischen Beispielen verdeutlichen.

Der erste Funktionskreis ist auf die Frage bezogen: Wie und mit welchen (digitalen) Mitteln kann man wissenschaftliche Forschungsliteratur (in digitaler Form) für eine spezifische Fragestellung ermitteln und beschaffen? Die Einklammerung des digitalen Faktors soll hier andeuten, dass weder die digitale Suche noch das digitale Format von gefundenen Resultaten ein Wert an sich ist. Die Erfahrung lehrt aber, dass der erste und primäre Zugang zur Suche nach Forschungsliteratur und nach Quellen heute digitale Findmittel sind. Und auch im Hinblick auf die Resultate haben manche Nutzer heute eine Vorliebe für digital verfügbare, unkompliziert reproduzierbare und leicht „mitnehmbare" Ressourcen. Dieser Funktionskreis wird in erster Linie „bedient" durch die digitalen Kataloge der Bibliotheken, die heute nicht nur selbstständige Veröffentlichungen, sondern auch Zeitschriftenartikel und Buchkapitel erschließen, sodann durch fachliche Portale, digitale Versionen von Fachbibliographien,[38] bibliographische Newsletter zu bestimmten Themengebieten und durch kommerzielle Angebote wie Google Books, Amazon usw., die wiederum eng mit den Resultaten von Suchmaschinen zusammenspielen. Man könnte meinen, diese basale Funktion ließe sich „abhaken". Es bleibt aber das Nadelöhr der Verschlagwortung und der „Übersetzung" von Forschungsinteressen in Abfragekombinationen. Wenn man sich zum Beispiel für die sprachlichen Verfahren der Gestaltung von literarischen Figuren interessiert: Welche Deskriptoren sind einschlägig? Wenn man sich für Lehrbücher als Gegenstände linguistischer Forschung interessiert: Wie kann man die Masse der Lehrbücher trennen von den Studien, die Lehrbücher untersuchen? So bleiben derzeit, auch jenseits der Anleitungen zum Bibliographieren, Unsicherheiten und Unwägbarkeiten, denen sich beispielsweise interdisziplinäre Ansätze zwischen Informationswisssenschaft und Informatik zum „Semantic Web" widmen.

38 Z. B. Bibliographie der deutschen Sprach- und Literaturwissenschaft: http://www.bdsl-online.de; Bibliography of Linguistic Literature: http://www.blldb-online.de; Angebote der Bibliothek des Instituts für deutsche Sprache: http://www1.ids-mannheim.de/bibliothek (16. 1. 2018).

Der zweite Funktionskreis bezieht sich auf Quellen und Sprachdaten für eine spezifische Fragestellung: Wie und mit welchen Mitteln kann man relevante Quellen und textuelle oder mediale Sprachdaten und ggf. multimodale Daten für die Bearbeitung einer wissenschaftlichen Fragestellung ermitteln und beschaffen? Im Hinblick auf literarische Quellen stellen zunächst digitale Bibliothekskataloge das Mittel der Wahl dar. Im Hinblick auf Sprachdaten unterschiedlicher Art ist das „Virtual Language Observatory“[39] (VLO) auch für deutschsprachige Angebote ein reichhaltiges Verzeichnis, bei dem sich Suchfacetten wie „Language = German“ kombinieren lassen mit Suchwörtern wie z. B. „teacher“, die dann etwa zu Gesprächsaufnahmen im „Forschungs- und Lehrkorpus Gesprochenes Deutsch“[40] des Instituts für Deutsche Sprache führen, in denen eine der Gesprächsrollen ein Lehrer oder eine Lehrerin ist. Kombiniert man „Language = German“ mit einem Suchwort wie „Frauenstimmrecht“, dann gelangt man zu entsprechenden Texten aus der ersten Frauenbewegung, die im „Deutschen Textarchiv“[41] der Berlin-Brandenburgischen Akademie der Wissenschaften angeboten werden. Das VLO ist also ein Werkzeug, mit dem sich digitale Daten und Tools ganz unterschiedlicher Anbieter mit Hilfe von Suchfacetten ermitteln lassen, die auf reichhaltige und komplex strukturierte Metadaten bei der Verschlagwortung der Ressourcen zugreifen.

Ein dritter Funktionskreis bezieht sich auf die Frage, ob es für eine wissenschaftliche Problemstellung bereits (etablierte, bewährte) digitale Bearbeitungsszenarien gibt und wie sie ggf. dokumentiert sind. Und wenn es entsprechende Bearbeitungsszenarien gibt: Was sind und wie findet man ggf. taugliche digitale Daten und digitale Werkzeuge, was sind bewährte Formen der koordinierten Nutzung solcher Ressourcen für eine bestimmte wissenschaftliche Fragestellung? Nun könnte man einerseits sagen, dass jede Problemstellung eigene methodische Entscheidungen erfordert. Auf der anderen Seite gibt es aber z. B. bei der sprachwissenschaftlichen Materialauswertung und vielfach auch im heuristischen Vorfeld von Untersuchungen solche (elementaren) Fragen wie: Wo und wie kann man ermitteln, welche Wortbildungen im Umkreis von „Flucht“ im Jahr 2017 im öffentlichen Diskurs verwendet wurden? Oder: Welche Rolle spielen Partizipial-Attribute in Zeitungstexten? Wie kann man in großen Korpora überhaupt systematisch nach grammatisch-syntaktischen Strukturen suchen? Für solche Aufgaben gibt es im Rahmen von

---

39 https://vlo.clarin.eu (16. 1. 2018) sowie Van Uytvanck, Stehouwer & Lampen (2012).

40 http://agd.ids-mannheim.de/folk.shtml (16. 1. 2018).

41 http://www.deutschestextarchiv.de (16. 1. 2018). Im VLO werden Ressourcen mit stabilen Adressen angegeben, z. B. http://hdl.handle.net/11858/00-203Z-0000-002E-72FE-4 für einen der Texte zur Frauenstimmrechtsbewegung im Deutschen Textarchiv (16. 1. 2018).

Korpusanfragesprachen durchaus bewährte Routinen und Vorgehensweisen. Sie sind nach meiner Einschätzung bislang aber noch nicht breit genug dokumentiert.[42] Es gehört deshalb mit zu den Aufgaben der nächsten Jahre, wissenschaftliche Nutzungsszenarien für digitale Daten und Werkzeuge breit und in Bezug auf unterschiedliche Einsatzbereiche in Lehre und Forschung zu beschreiben (dazu unten mehr).

4 Infrastrukturverbünde und Nutzereinbindung: Perspektiven und Aufgaben aus einer germanistischen Sicht

Welche Perspektiven und Zukunftsaufgaben ergeben sich im Bereich der Germanistik für den weiteren Ausbau von Infrastrukturverbünden wie CLARIN-D und DARIAH-DE? Aus einer Nutzerperspektive[43] kann man folgende wichtige Möglichkeiten sehen, wie Wissenschaftler/-innen sich an der weiteren Gestaltung und dem weiteren Ausbau von Infrastrukturangeboten beteiligen können:

a) Beitrag zur Bedarfsanalyse: Welche (Arten von) Daten und Werkzeugen fehlen im Bestand oder sollten ausgebaut werden?
b) Beitrag zur Bewertung von vorhandenen Angeboten im Hinblick auf die Bedürfnisse von fachlichen Nutzergruppen, hierzu gehören z. B. auch Fragen der Gebrauchstauglichkeit (usability)[44] von Angeboten oder die Frage, wie Ressourcen auf wissenschaftliche Fragestellungen zu beziehen sind.
c) Beitrag zur Ressourcendokumentation und -akquise: Wo gibt es nützliche Ressourcen (Daten, Werkzeuge), die sich sinnvollerweise in eine Infrastruktur eingliedern lassen? Wie kann man Kolleg/-innen ermuntern,

---

42 Mehrere Beispiele für unterschiedliche Dokumentationsformen: https://www.youtube.com/user/CLARINGermany (16. 1. 2018).

43 Andere Perspektiven sind etwa die Fragen der technischen Einrichtung von Infrastrukturmaßnahmen, der Aspekt der dauerhaften Finanzierung und der Verstetigung von Angeboten oder auch die Frage der Koordination und der Abstimmung verschiedener Initiativen, die in unterschiedlicher Weise zur digitalen Arbeitslandschaft im Bereich der Germanistik beitragen. – Diese Themen und Fragestellungen sind und waren in vielen Arten von Infrastrukturen immer schon virulent, vgl. Edwards et al. (2009).

44 Die Usability-Forschung ist inzwischen nicht mehr zu überblicken. Zeitschriften wie das „Journal for Usability Studies“ enthalten immer wieder auch Beispiele für Usability-Untersuchungen zu digitalen Angeboten, z. B. Brown & Hocutt (2015). Aber auch Einzelstudien können Anregungen zur Methodik geben, z. B. Schulz (2016).

Daten und Werkzeuge für die nachhaltige Nutzung im Rahmen einer Infrastruktur bereitzustellen?

d) Beitrag zur Dissemination von Angeboten: Wie und mit welchen Mitteln kann man digitale Angebote und ihre Nutzungsweisen in den unterschiedlichen Fachzonen der Germanistik[45] bekannt(er) machen?

e) Wie lassen sich fachlich nützliche Angebote, die nicht Bestandteil von Infrastrukturen sind, sinnvoll mit den Infrastrukturangeboten verbinden oder immerhin für die Nutzergruppen dokumentieren?[46]

f) Beitrag zur zielgruppenorientierten Dokumentation von typischen Anwendungsweisen digitaler Ressourcen (Daten, Werkzeuge) für wissenschaftliche Problemstellungen. Zu den Zielgruppen gehören nicht nur Forschende auf allen Qualifikationsstufen, sondern auch die Studierenden.

Den Aufgabenbereich (f) halte ich für besonders wesentlich, denn er hängt mit dem innersten funktionalen Kern von Infrastrukturen zusammen: Die Infrastrukturen sollen dazu beitragen, dass Wissenschaftler und Wissenschaftlerinnen ihre Aufgaben in Forschung und Lehre gut erfüllen und digitale Daten und Werkzeuge hierfür in sachdienlicher Weise einsetzen können. Die Dokumentation von Nutzungsszenarien ist eine Gelenkstelle, die das Infrastruktur-Angebot mit seinen Nutzungsmöglichkeiten verbindet.

Für diese Zukunftsaufgaben sind im Bereich der Germanistik drei wichtige Bezugspunkte und Rahmenbedingungen zu berücksichtigen, die ich hier zunächst nenne und im Folgenden genauer erläutere: zum einen (i) die äußerst breite fachliche Differenzierung in der Germanistik, zum zweiten (ii) die im Hinblick auf Voraussetzungen und Interessen sehr unterschiedlichen Nutzergruppen und schließlich (iii) die Frage, wie und mit welchen Informationsmitteln und Darstellungsstrategien man diese heterogene(n) Nutzergruppe(n) erreichen kann.

---

45 Eine interessante Frage ist dabei auch, wann, wie, in welcher Weise und in welchem Umfang die Nutzung digitaler Daten und Ressourcen in das Selbstverständnis einer fachlichen Gemeinschaft „hineinwächst". Vgl. z. B. die Beiträge von Maitz, Nübling und Bubenhofer & Scharloth in Maitz (2012).

46 Bei einem gemeinsamen Workshop des DFG-Netzwerks „Diskurse digital" und der CLARIN-D-Facharbeitsgruppe „Deutsche Philologie" im Dezember 2017 haben u. a. Noah Bubenhofer und Simon Meier-Vieracker darauf hingewiesen, dass vor allem in der Forschung zu aktuellem Social Media-Sprachgebrauch auch Techniken benötigt werden, die tagesaktuell angewendet werden können und die bislang nicht im Bestand von Infrastrukturangeboten sind, z. B. die Nutzung der APIs von Angeboten wie Twitter, Facebook oder YouTube. Werkzeuge dafür gibt es etwa im Umkreis der R-Bibliothek. Der schon ältere Gegensatz zwischen der Infrastrukturstrategie und eher offenen, dynamischen „digital ecosystems" wird in einem Beitrag von Blanke, Kristel & Romary (2016) beleuchtet.

(i) Die Fachzone der Germanistik ist durch eine äußerst breite fachliche Differenzierung gekennzeichnet, deren Teile in manchen Hinsichten verbunden, in anderen Hinsichten aber fachlich sehr weit voneinander entfernt sind.[47] Sprache und Literatur, die darauf bezogenen Didaktiken, gegenwartssprachliche und historische Bezüge, die unterschiedlichen Beschreibungsebenen von der Phonetik bis zur Pragmatik, die unterschiedlichen Varietäten des Deutschen von den Dialekten über Fachsprachen bis hin zu Gruppensprachen wie dem Kiezdeutsch und viele andere gegenstandsbezogene, theoretische und methodisch motivierte Unterscheidungen tragen zur Vielfalt der germanistischen Forschung und ihrer Lehrgebiete bei. Wenn jemand den Fremdwortanteil in den Werken von Simon Dach (17. Jh.) untersucht, ist das ein grundsätzlich anderer Untersuchungsbereich als etwa eine Untersuchung zu bestimmten syntaktischen Besonderheiten in ausgewählten Dialekten des Deutschen oder zu vielen literaturwissenschaftlichen Themen, obwohl alle diese Themen unterm Dach der Germanistik beheimatet sind. Sowohl im Hinblick auf Bedarfsanalysen als auch im Hinblick auf die Dokumentation von Nutzungsszenarien für digitale Daten und Werkzeuge empfiehlt sich eine hohe Granularität, also die Berücksichtigung kleiner Fachzonen innerhalb der großen Germanistik.

(ii) Neben der Vielfalt in der fachlichen Organisation muss man eine weitere Differenzierung im Hinblick auf Nutzergruppen mit unterschiedlichen Zielen und Voraussetzungen berücksichtigen. Auch in der Germanistik gibt es eine sehr aktive Gruppe, die man als „Digital Humanities-Avantgarde“ bezeichnen könnte. Diese Avantgarde kümmert sich unter anderem darum, den Bereich der Möglichkeiten, der durch digitale Daten und Werkzeuge eröffnet wird, im Hinblick auf neue und neuartige Forschungsperspektiven zu erweitern. Demgegenüber gibt es im Bereich der Germanistik auch den „Alltagsbetrieb“ einer digital unterstützten Germanistik. Dazwischen gibt es Ansätze, die darauf zielen, auch traditionelle disziplinäre Fragestellungen mit digitalen Daten und Werkzeugen anders und ggf. besser zu bearbeiten. Zwischen den beiden Polen der Avantgarde und der Alltagsforschung gibt es einen breiten Übergangsbereich. Viele germanistische Konferenzvorträge, die ich gehört habe, behandelten im Kern traditionell motivierte Themen, die aber mit Hilfe digitaler Daten und Werkzeuge, etwa des „Deutschen Textarchiv“ oder der IDS-Korpora, auf durchaus innovative Weise bearbeitet wurden. Solche Arbeiten sind eine gute Grundlage, um die (erfolgreiche) computergestützte Untersuchung wissenschaftlicher Fragestellungen im Zusammenhang von typisierten Nutzungsszenarien zu dokumentieren.

47 Die (fehlende) Einheit des Faches „Germanistik“ ist seit vielen Jahren ein Thema der innerfachlichen Diskussion.

(iii) Damit stellt sich auch die dritte Frage, wie man typisierte Nutzungsszenarien für die unterschiedlichen Nutzergruppen am besten dokumentiert. Im Rahmen der CLARIN-D-Facharbeitsgruppe „Deutsche Philologie“ haben wir unter anderem mit Schritt-für-Schritt-Anleitungen, mit Video-Screencasts und mit sog. Video-Experten-Interviews gute Erfahrungen gemacht. Dabei folgen wir jeweils einem thematischen Viererschema: Wir gehen (1) von einer wissenschaftlichen Fragestellung oder einem Typ von Fragestellung aus, die bzw. der kurz vorgestellt wird, wir stellen sodann (2) die für die Bearbeitung genutzten Daten und Werkzeuge vor und erläutern, wo man sie finden kann, wir beschreiben dann (3) die koordinierte Anwendung der Werkzeuge und der Daten im Hinblick auf die Fragestellung (ggf. in einer Schritt-für-Schritt-Beschreibung), zuletzt (4) stellen wir die im Beispiel erzielten Resultate dar und beziehen sie zurück auf die Ausgangsfragestellung. Wenn es bereits Publikationen gibt, die auf dem besprochenen Nutzungsszenario beruhen, nennen wir diese Publikationen ebenfalls. Auf dem oben erwähnten CLARIN-D-YouTube-Kanal sind Beispiele für Screencasts und für den etwas aufwändigeren Typ des Experten-Interviews zu sehen.

5 Rückblick

Mit der Digitalisierung sind erhebliche Veränderungen auch in der wissenschaftlichen Arbeitslandschaft von Germanist/-innen in den ganz unterschiedlichen germanistischen Fachzonen verbunden. In diesem Beitrag habe ich versucht, einige wesentliche Veränderungen zu skizzieren, die sich aus der Verfügbarkeit digitaler Daten und Werkzeuge für den Alltag von Wissenschaftler/-innen, insbesondere ihre Forschungs- und Lehrpraxis ergeben. Im Fokus stand sodann die Frage, wie Infrastrukturangebote zur produktiven Gestaltung der Arbeitslandschaft und der Forschungsmöglichkeiten beitragen können und wie Nutzergruppen bzw. -vertretungen dazu beitragen können, Infrastrukturangebote mit den wissenschaftlichen Bedürfnissen ihrer Nutzer/-innen abzustimmen.

Literatur

Bader, Anita (2018): *Mailinglists als Format der digitalen Wissenschaftskommunikation. Eine linguistische Untersuchung*. Diss. Universität Gießen. Gießen: Gießener Elektronische Bibliothek. http://geb.uni-giessen.de/geb/volltexte/2018/13523 (letzter Zugriff: 28. 5. 2018).

Bender, Michael (2016): *Forschungsumgebungen in den Digital Humanities. Nutzerbedarf, Wissenstransfer, Textualität*. Berlin/Boston: De Gruyter.

Blanke, Tobias, Conny Kristel & Laurent Romary (2016): Crowds for clouds. Recent trends in humanities infrastructures. In Agiati Bernardou et al. (Hrsg.), *Cultural heritage digital tools and Infrastructures*. https://hal.inria.fr/hal-01248562 (letzter Zugriff: 20. 1. 2018).

Brown, David J. (2016): *Access to scientific research. Challenges facing communication in STM*. Berlin/Boston: De Gruyter.

Brown, Maury Elizabeth & Daniel L. Hocutt (2015): Learning to use, useful for learning: A usability Study of Google apps for education. *Journal of Usability Studies* 10 (4), 160–181.

Deegan, Marilyn & Willard McCarty (Hrsg.) (2012): *Collaborative research in the Digital Humanities*. Surrey, UK/Burlington, USA: Ashgate.

Edwards, Paul N. et al. (2009): Introduction: An agenda for infrastructure studies. *Journal of the Association for Information Systems* 10 (5), 364–374.

Flanders, Julia (2009): The productive unease of 21st-century digital scholarship. *Digital Humanities Quarterly* 3 (3). http://digitalhumanities.org/dhq/vol/3/3/000055/000055.html (letzter Zugriff: 20. 1. 2018).

Gloning, Thomas & Gerd Fritz (Hrsg.) (2011): *Digitale Wissenschaftskommunikation. Formate und ihre Nutzung*. Gießen: Gießener Elektronische Bibliothek. http://geb.uni-giessen.de/geb/volltexte/2011/8227 (letzter Zugriff: 20. 1. 2018).

Habel, Thomas (2007): *Gelehrte Journale und Zeitungen der Aufklärung. Zur Entstehung, Entwicklung und Erschließung deutschsprachiger Rezensionszeitschriften des 18. Jahrhunderts*. Bremen: edition lumière.

Hanisch, Ludmila (Hrsg.) (2000): *„Machen Sie doch unseren Islam nicht gar zu schlecht". Der Briefwechsel des Islamwissenschaftlers Ignaz Goldziher und Martin Hartmann 1894–1914*. Hg. und kommentiert. Wiesbaden: Harrassowitz.

Krappmann, Lothar (1993): Gespräche über das Computernetz? In Gerold Becker & Jürgen Zimmer (Hrsg.): *Lust und Last der Aufklärung. Ein Buch zum 80. Geburtstag von Hellmut Becker*, 31–43. Weinheim/Basel: Beltz.

Lobin, Henning (2009): *Inszeniertes Reden auf der Medienbühne. Zur Linguistik und Rhetorik der wissenschaftlichen Präsentation*. Frankfurt a. M.: Campus.

Lobin, Henning (2014): *Engelbarts Traum. Wie der Computer uns Lesen und Schreiben abnimmt*. Frankfurt/New York: Campus.

Maitz, Péter (Hrsg.) (2012): *Historische Sprachwissenschaft. Erkenntnisinteressen, Grundlagenprobleme, Desiderate*. Berlin/Boston: de Gruyter.

McCarty, Willard (2012): Collaborative research in the Digital Humanities. In: Marilyn Deegan & Willard McCarty (Hrsg.): *Collaborative research in the Digital Humanities*, 1–10. Surrey, UK/Burlington, USA: Ashgate.

Moretti, Franco (2007): *Graphs, maps, trees. Abstract models for literary history*. London/New York: Verso.

Moretti, Franco (2013): *„Operationalizing": or, the function of measurement in modern literatary Theory*. Stanford Literary Lab, Pamphlet 6, December 2013. https://litlab.stanford.edu/LiteraryLabPamphlet6.pdf (letzter Zugriff 12. 11. 2017)

Sahle, Patrick (2013): *Digitale Editionsformen*. Drei Bände. Norderstedt: BoD. https://www.i-d-e.de/publikationen/schriften/s7-9-digitale-editionsformen (letzter Zugriff 20. 1. 2018).

Schneider, Roman & Horst Schwinn (2014): Hypertext, Wissensnetz und Datenbank: Die Web-Informationssysteme grammis und ProGr@mm. In: *Institut für Deutsche*

Sprache: Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache, 337–346. Mannheim: IDS.

Schnettler, Bernt & Hubert Knoblauch (Hrsg.) (2007): *Powerpoint-Präsentationen. Neue Formen der gesellschaftlichen Konstruktion von Wissen*. Konstanz: UVK.

Schulz, Arne Hendrik (2016): *Usability in digitalen Kooperationsnetzwerken. Nutzertests und Logfile-Analyse als kombinierte Methode*. Dissertation Uni Bremen. Bremen: UB. urn:nbn:de:gbv:46-00105045-14 (letzter Zugriff: 20. 1. 2018).

Van Uytvanck, Dieter, Herman Stehouwer & Lari Lampen (2012): Semantic metadata mapping in practice: The Virtual Language Observatory. In Nicoletta Calzolari (Hrsg.): *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC 2012), Istanbul, May 23rd–25th 2012, 1029–1034.

Weinrich, Harald (1995a): Sprache und Wissenschaft. In: Heinz L. Kretzenbacher & Harald Weinrich (Hrsg.): *Linguistik der Wissenschaftssprache*, 3–13. Berlin/New York: de Gruyter.

Weinrich, Harald (1995b): Wissenschaftssprache, Sprachkultur und die Einheit der Wissenschaft. In: Heinz L. Kretzenbacher & Harald Weinrich (Hrsg.): *Linguistik der Wissenschaftssprache*, 155–174. Berlin/New York: de Gruyter.

Erhard Hinrichs

2 Digitale Forschungsinfrastrukturen für die Sprachwissenschaft

Abstract: Linguistische Forschungsdaten zeichnen sich durch eine große Vielfalt unterschiedlicher Datentypen aus, die sich aus methodischen Grundannahmen des Faches, aber auch aus den verschiedenen Teildisziplinen der Sprachwissenschaft ergeben. Digital verfügbare Daten ermöglichen neue Formen des Datenzugangs und erfordern neue Werkzeuge im Umgang mit Daten, speziell zur Datenrecherche und zur Datenanalyse. Daraus ergeben sich umfängliche fachspezifische Anforderungen an eine digitale Forschungsinfrastruktur für die Sprachwissenschaft. Diese Anforderung und deren mögliche Lösungen werden exemplarisch anhand der CLARIN-Forschungsinfrastruktur für die Geistes- und Sozialwissenschaften erläutert.

Keywords: CLARIN, Forschungsdaten, Interoperabilität, vernetze Forschungsinfrastruktur

Anmerkung: Als verteilte Infrastrukturinitiative sind viele Einzelpersonen an CLARIN beteiligt, die an den CLARIN-Zentren die Infrastruktur betreiben, Ressourcen bearbeiten und bereitstellen, Expertise und Beratungsleistungen einbringen. Ohne sie wäre CLARIN und dieser Artikel nicht möglich. Ich danke auch den Forschenden, die in den CLARIN-D-Facharbeitsgruppen mitarbeiten und wertvolle Rückmeldungen zu den Werkzeugen und Diensten von CLARIN-D geben. Sie initiieren an den Bedarfen der Fachdisziplinen orientierte Weiterentwicklungen und geben als Berater/-innen und als Testanwender/-innen wichtige Hinweise bei der Entwicklung und Nutzung von Ressourcen, virtuellen Forschungsumgebungen und Diensten.

Die CLARIN-Zentren werden in Deutschland seit 2008 durch das Bundesministerium für Bildung und Forschung (BMBF) gefördert; der Projektträger ist das Deutsche Zentrum für Luft- und Raumfahrt (DLR). Außerdem beteiligen sich die Institutionen, die CLARIN-D-Zentren beherbergen, und verschiedene Bundesländer an der Finanzierung. Diese föderale Förderung hat wesentlich zur Entwicklung von Forschungsinfrastrukturen für die Geistes- und Sozialwissenschaften in Deutschland und Europa beigetragen und hat auch diese Publikation ermöglicht.

Es ist mir ein besonderes Anliegen, Herrn Ludwig M. Eichinger, meinem Kollegen und Mitstreiter in CLARIN-D und langjährigem Direktor des Instituts für Deutsche Sprache in Mannheim, für viele Jahre guter Zusammenarbeit zu danken. Mit seiner Weitsicht, Dateninfrastrukturen am Institut für Deutsche Sprache nachhaltig zu etablieren und gezielt auszubauen, hat er wesentlich zum Aufbau und zur Entwicklung von Forschungsinfrastrukturen für die Sprachwissenschaft in Deutschland beigetragen.

Erhard Hinrichs, Seminar für Sprachwissenschaft, Eberhard Karls Universität Tübingen, Wilhelmstr. 19, D-72074 Tübingen, E-Mail: erhard.hinrichs@uni-tuebingen.de

https://doi.org/10.1515/9783110538663-003

1 Einleitung

Die Sprachwissenschaft hat im 20. Jahrhundert eine rasante Entwicklung hinsichtlich ihrer theoretischen Durchdringung und der Breite ihrer Anwendungsgebiete durchlaufen. Aufbauend auf den bahnbrechenden Arbeiten der Junggrammatiker in der zweiten Hälfte des 19. Jahrhunderts, die sich auf Fragen der diachronen Sprachwissenschaft, speziell des Lautwandels, konzentrierten und zur Rekonstruktion früherer Sprachzustände („Protosprachen") die historischvergleichende Methode entwickelte, nahm der europäische (de Saussure 1916) und amerikanische (Bloomfield 1914; Sapir 1921) Strukturalismus mit Beginn des 20. Jahrhunderts die synchrone Sprachwissenschaft und die Erforschung von grammatischen Systemen insgesamt in den Blick.

In der zweiten Hälfte des 20. Jahrhunderts, entwickelte sich die generative Transformationsgrammatik zu einem der führenden Paradigmen der Sprachwissenschaft, dessen biologische und kognitionswissenschaftliche Grundannahmen zu einer Universalgrammatik sich nicht nur auf die Erforschung von Einzelsprachen auswirkte, sondern auch neue Forschungsmethoden in verschiedenen Teilgebieten der Sprachwissenschaft stimulierte, wie etwa zum Erst- und Zweitsprachenerwerb, zur Sprachverarbeitung und zur mathematischen Modellierung von Sprache.

Gleichzeitig rückte das generative Programm Chomskys (Chomsky 1957) mit intuitiven Sprecherurteilen von Muttersprachler/-innen einen neuen Datentyp als empirische Grundlage für linguistische Analysen in den Vordergrund. Die sich daraus ergebenden Schwierigkeiten und erforderlichen methodischen Standards bei der Datenerhebung rückten erst allmählich in das Bewusstsein der Fachdisziplin (Schütze 1996), die in Deutschland u. a. durch den Tübinger Sonderforschungsbereich zum Thema *Linguistische Datenstrukturen: Theoretische und empirische Grundlagen der Grammatikforschung* (1999–2009) thematisiert wurden und zur Etablierung einer Fachkonferenz mit dem Titel *Linguistic Evidence* führten.[1]

Parallel zu dieser Methodenreflexion in der theoretischen Sprachwissenschaft haben seit den 1980er Jahren in vielen Teildisziplinen der Sprachwissenschaft experimentelle und statistische Verfahren zunehmend an Bedeutung gewonnen. Wichtige Impulse bei dieser Entwicklung gingen dabei von der

---

1 Die Bemühungen des Tübinger SFB 441 um eine nachhaltige Datenmodellierung und -archivierung führten zu einer neuen Entwicklung beim Förderformat von Sonderforschungsbereichen, in denen nun INF Teilprojekte als Instrument für ein nachhaltiges Management von Forschungsdaten etabliert wurden. Das INF Teilprojekt im Tübinger SFB 441 hatte in dieser Hinsicht Modellcharakter für die DFG und war das erste Teilprojekt dieser Art in einem DFG geförderten Sonderforschungsbereich.

Computerlinguistik (Klavans & Resnik 1997), der Psycholinguistik und der Phonetik und Phonologie, aber in den letzten Jahren auch von der Syntax (Bresnan 2016; Sprouse & Hornstein 2012) und von der Semantik und Pragmatik (Noveck & Sperber 2004; Meibauer & Steinbach 2011) aus. In der Computerlinguistik wurden statistische Verfahren zuerst in der automatischen Spracherkennung (Jelinek 1997), dann aber auch bei der automatischen Analyse geschriebener Sprache, etwa beim Wortartentagging (Church 1988; Schmid 1994; Brants 2000), beim Parsing (Collins 2003) und bei der maschinellen Übersetzung (Brown et al. 1990) eingesetzt. Inzwischen sind statistische Ansätze und maschinelle Lernverfahren in der Computerlinguistik zu etablierten Verfahren geworden. In der Psycholinguistik haben sich eine breite Palette experimenteller Verfahren wie Reaktionszeit- und Augenbewegungsmessungen, neurophysiologische Untersuchungsmethoden wie Elektroenzephalogrammen (EEG) und Ereigniskorrelierte hirnelektrische Potentiale (EKPs) sowie bildgebende Verfahren mit Hilfe von Positronenemissions-Tomographie (PET) und von funktioneller Kernspinresonanz-Tomographie (fMRI) etabliert und zu einer interdisziplinären Zusammenarbeit zwischen der Linguistik, der Kognitionspsychologie und den Neurowissenschaften geführt. Experimentelle Methoden in der Sprachwissenschaft haben sich ebenfalls auf der Schnittstelle von Signalverarbeitung, Phonetik und Phonologie unter dem Begriff *Laboratory Phonology* (Pierrehumbert, Beckman & Ladd 2000) als eine neuer methodischer Ansatz zur Erforschung von Lautsystemen unter Berücksichtigung ihrer physikalischen und physiologischen Basis etabliert.

2 Linguistische Forschungsdaten und digitale Sprachressourcen

Der Einsatz experimenteller und statistischer Methoden hat die empirische Fundierung linguistischer Forschung und den Stellenwert linguistischer Daten neu in den Blick genommen und zu einer großen Vielfalt von Forschungsdaten[2] geführt.

2 In Abhängigkeit von einzelnen Fachdisziplinen oder Wissenschaftszweigen bezeichnet der Begriff *Forschungsdaten* ganz unterschiedliche Datentypen. Es ist das Verdienst der Rates für Informationsinfrastrukturen (RfII), sich um eine fächerübergreifende Begriffklärung bemüht zu haben. (Rat für Informationsinfrastrukturen (RfII) 2014) definiert Forschungsdaten als „… Daten, die im Zuge wissenschaftlicher Vorhaben entstehen, z. B. durch Beobachtungen, Experimente, Simulationsrechnungen, Erhebungen, Befragungen, Quellenforschungen, Aufzeichnungen, Digitalisierung, Auswertungen.“

Forschungsdaten, d. h. Daten, die für wissenschaftliche Untersuchungen erhoben werden bzw. als Resultat solcher Untersuchungen entstehen, haben traditionell für die sprachwissenschaftliche Forschung eine wichtige Rolle gespielt. Beispielhaft genannt seien hier groß angelegte Datenerhebungen zur Dialektologie, wie sie von Wenker zwischen 1876 und 1888 für die Erfassung von Dialekten des Deutschen erhoben wurden und bis heute im Deutschen Sprachatlas der Universität Marburg archiviert werden. Wurden Forschungsdaten traditionell als Printmaterialien veröffentlicht und in Archiven oder Bibliotheken den jeweiligen Fachwissenschaften zur Verfügung gestellt, eröffnen sich durch die digitalen Medien neue Arten der Datenerhebung und der Datenzugänge, erfordern aber gleichzeitig neue Kompetenzen beim Umgang mit Daten und neue Wege bei der nachhaltigen Datenarchivierung. Auf diesem Hintergrund ist bei der Erstellung von digitalen Forschungsdaten ein umsichtiges Datenmanagement im Vorfeld eines Forschungsvorhabens von großer Bedeutung. Daher erwarten inzwischen nationale und internationale Förderorganisationen in allen Wissenschaftszweigen einen Datenmanagementplan als obligatorischen Bestandteil eines Förderantrags, in dem Forschungsdaten generiert werden sollen. Zum Datenmanagement gehört die Beachtung von Standards und von in der Fachdisziplin bewährten Methoden und Praktiken bei der Datenenkodierung, bei der Erstellung von Metadaten und der Datenarchivierung ebenso wie, im Falle von Datenerhebungen mit Informant/-innen, die Wahrung von Persönlichkeitsrechten und anderen ethischen Fragen.

2.1 Linguistische Forschungsdaten

Linguistische Forschungsdaten zeichnen sich im Vergleich zu anderen Fachdisziplinen durch eine große Vielfalt unterschiedlicher Datentypen aus. Zu diesen Datentypen gehören Informantenbefragungen für Studien zur Grammatik und zur Sprachvariationsforschung; Korpusdaten, häufig angereichert durch linguistische Annotationen, als empirische Grundlage zum Sprachgebrauch oder als Trainingsmaterial für statistische Sprachmodelle und für überwachte maschinelle Lernverfahren in der Computerlinguistik; Sprachaufnahmen für phonetische und phonologische Studien sowie sprachliche Stimuli für experimentelle Verfahren in der Psycholinguistik. Werden experimentelle Analyseverfahren auf linguistische Forschungsprimärdaten angewendet und die dabei erzielten Ergebnisse ausgewertet, so entstehen dadurch neue sekundäre Forschungsdaten, die in weiteren Untersuchungen als Referenzmaterialien dienen können. Für eine solche Nachnutzung ist die Provenienz dieser Sekundärdaten von entscheidender Bedeutung. Datenprovenienz bedeutet, dass die eingesetzten Analyseverfahren, einschließlich der dabei eingesetzten

Software und den verwendeten Parametern, die zur Generierung der Sekundärdaten geführt haben, protokolliert werden und als technische Metadaten mit den Sekundärdaten verknüpft werden.

2.2 Digitale Sprachdaten

Die Digitalisierung und die rasante Entwicklung des Internets im Übergang vom 20. ins 21. Jahrhundert hat diese methodischen Entwicklungen zusätzlich verstärkt und beschleunigt. Durch die Digitalisierung liegen nun Sprachdaten in einer zuvor nie gekannten Größenordnung vor. Dadurch hat sich der Umfang von Referenzkorpora für Einzelsprachen in den letzten 25 Jahren dramatisch verändert. So hatte das British National Corpus (Burnard & Aston 1998), das zwischen 1991 und 1994 entstanden ist, einen Umfang von 100 Millionen Wörtern. Das deutsche Referenzkorpus (Kupietz et al. 2010) mit gegenwärtig 31,68 Milliarden Wörtern (Stand 8. 3. 2017), das am Institut für Deutsche Sprache entstanden ist und kontinuierlich erweitert wird, und sog. *Webkorpora*, d. h. Textsammlungen von Internetdaten, die durch Web-crawling-Verfahren zu digitalen Korpora in Milliardengröße zusammengeführt werden, zeigen diese Entwicklung exemplarisch auf. Gleichzeitig entstehen mit Sprachdaten aus sozialen Netzwerken und online Foren gänzlich neue Textsorten, die für die linguistische Forschung hohe Relevanz haben.

Mit Hilfe von Retrodigitalisierungsverfahren wie der optischen Zeichenerkennung lassen sich Druckerzeugnisse in digitale Formate von Bilddigitalisaten und von Volltexten überführen und der wissenschaftlichen Forschung zur Verfügung stellen. Retrodigitalisierungsverfahren werden in der Sprachwissenschaft vor allem zur Erforschung der Sprachvariation, des Sprachwandels und der Sprachevolution eingesetzt. Exemplarisch sei hier das Deutsche Textarchiv (DTA) (Geyken et al. 2011) genannt, ein umfangreiches historisches Korpus für das Neuhochdeutsche mit Texten ab 1700, das als Sammlung von Bilddigitalisaten und von linguistisch aufbereiteten Volltexten zur Verfügung steht. Das DTA enthält wissenschaftliche Texte, Gebrauchsliteratur und Belletristik und umfasst gegenwärtig 2642 Werke 613.312 digitalisierte Seiten und 146.891.533 fortlaufende Wortformen (Stand: Mai 2017). Die Texte sind mit Hilfe computerlinguistische Analyseverfahren auf der Wortebene in ihrer Orthografie normalisiert, lemmatisiert und mit morpho-syntaktischen Wortklassen annotiert. Diese linguistischen Annotationen ermöglichen es, das DTA für diachrone linguistische Untersuchungen zu nutzen. In meinen Studien zur historischen Syntax des Verbalkomplexes im Deutschen (Hinrichs 2016) und zur morphologischen Produktivität von Adjektiven des Deutschen aus diachroner Perspektive (Hinrichs 2017) nutze ich die DTA Daten und deren linguistische

Annotationen als Datengrundlage. In meinen Untersuchungen zur morphologischen Produktivität von Adjektiven habe ich außerdem von den umfänglichen Korpusstudien von (Eichinger 1982) zu Adjektivbildungen im Deutschen profitiert.[3]

Bei digital verfügbaren Sprachdaten, wie sie durch das Internet in großem Umfang zur Verfügung stehen, handelt es sich in vielen Fällen zunächst um unstrukturierte Daten. Um sie für die linguistische Forschung nutzbar zu machen, sind in der Regel umfangreiche Vorverarbeitungsschritte erforderlich. Dazu gehören die Erkennung von Satzgrenzen sowie die Tokenisierung, die Lemmatisierung, die Wortartenklassifikation und die morphologische Analyse von lexikalischen Einheiten. Mit computerlinguistischen Verfahren lassen sich diese Verarbeitungsschritten inzwischen mit hinreichend hoher Präzision realisieren. Allerdings arbeiten die Verfahren nicht vollkommen fehlerfrei, was bei der Benutzung derartig aufbereiteter Quellen stets berücksichtigt werden muss.

Die Verfügbarkeit digitaler Sprachdaten hat die empirischen Grundlagen der Sprachwissenschaft nicht nur in quantitativer Hinsicht verändert. Sie hat darüber hinaus neue Methoden der linguistischen Modell- und Theoriebildung ermöglicht. Die Lexikografie und die historische Sprachwissenschaft seien hier exemplarisch erwähnt. Fragen der Sprachevolution lassen sich auf der Grundlage digitaler Datensätze mit Hilfe von bioinformatischen Verfahren der Phylogenese untersuchen (Gray, Drummond & Greenhill 2009). In der Lexikografie können Lexikoneinträge mit Belegstellen aus digitalen Korpora verknüpft werden, um authentische Verwendungsbeispiele und -kontexte von Wörtern als Wortprofile (Kilgarriff et al. 2004; Geyken, Didakowski & Siebert 2008) online zur Verfügung zu stellen. Dieser Ansatz wird im Digitalen Wörterbuch der deutschen Sprache (DWDS) auf der Grundlage eines umfangreichen, balancierten Textkorpus umgesetzt (Klein & Geyken 2010).

Ein weiterer innovativer Ansatz in der Computerlexikografie besteht in der Modellierung von lexikalischer Semantik mit Hilfe von Graphstrukturen, wie sie im Princeton WordNet (Miller et al. 1990) und im FrameNet (Baker, Fillmore & Lowe 1998) für das Englische vorbildhaft umgesetzt worden sind. Analoge digitale lexikalische Ressourcen stehen für zahlreiche weitere Sprachen zur Verfügung, u. a. für das Deutsche mit den Ressourcen GermaNet (Henrich & Hinrichs 2010) und der FrameNet Ressource Salsa (Erk et al. 2003).

3 Ludwig M. Eichinger stützt sich in seinen Studien zu den Adjektiven des Deutschen bereits in den 1970er Jahren auf digitale Textkorpora. Er gehört damit zu den Pionieren in der germanistischen Linguistik und in der Allgemeinen Sprachwissenschaft im deutschsprachigen Raum, die in ihrer Forschung digitale Korpusdaten verwenden.

3 Anforderungen an eine Forschungsinfrastruktur für die Sprachwissenschaft und für sprachbasierte Forschung in den digitalen Geisteswissenschaften

Digital verfügbare Daten ermöglichen im Gegensatz zu traditionellen Druckerzeugnissen neue Formen des Datenzugangs und erfordern daher auch im Wissenschaftskontext neue Werkzeuge im Umgang mit Daten, speziell zur Datenrecherche und Datenanalyse sowie zur Visualisierung von Rechercheergebnissen. Derartige Softwarewerkzeuge sind in der Regel disziplinenspezifisch und erfordern fachnahe Entwicklungskompetenz. Sie sind außerdem eng mit den zugrundeliegenden digitalen (Forschungs-)Daten verknüpft, um einen gemeinsamen Zugang zu den Daten und zugehörigen Softwarewerkzeugen zu ermöglichen. Um den Zugang zu Daten und Werkzeugen möglichst einfach zu gestalten, werden solche digitale Angebote zunehmend als Webapplikationen realisiert, deren Zugang bei lizenzrechtlich geschützten Daten oder Werkzeugen durch ein Authentifizierungs- und Autorisierungsprotokoll unterstützt wird. Dies trifft in besonderer Weise auf Sprachressourcen zu, die im Gegensatz zu vielen anderen Typen von Forschungsdaten in der Regel nicht frei von Rechten Dritter sind. Betroffen sind insbesondere Urheber- und verwandte Schutzrechte sowie Persönlichkeitsrechte. Erschwerend kommt hinzu, dass die Rechteinhaber i. d. R. nicht Teil der Fachdisziplinen sind. Bei der Nutzung digitaler Sprachdaten sind daher urheberrechtliche Fragen stets zu berücksichtigen. Das betrifft insbesondere Datenquellen, die dem Copyright unterliegen und daher nicht frei weitergegeben werden dürfen. In der wissenschaftlichen Praxis führt dies häufig zu lizenzrechtlichen Zugangsbeschränkungen und erfordert elektronische Identifizierungs- und Autorisierungsverfahren.

Der digitale Zugang zu Forschungsdaten und Forschungssoftware ermöglicht neben der Einzelforschung das kollaborative Forschen und das Teilen von Forschungsdaten, das durch entsprechende Kommunikationssoftware und virtuelle Speicherlösungen unterstützt wird und zur Entwicklung von virtuellen Forschungsumgebungen geführt hat. Virtuelle Forschungsumgebungen bezeichnen innovative Forschungsplattformen, in denen einschlägige Datenressourcen, Verarbeitungs- und Visualisierungswerkzeuge sowie Kommunikations- und Speicherlösungen für das kollaborative Forschen und für das Teilen von Forschungsdaten, häufig netzbasiert, institutionsübergreifend und gff. auch über Fächer- und Ländergrenzen hinaus bereitgestellt werden.

Neben den neuen Formen des Zugangs zu digitalen (Forschungs-)daten sind neue Fachkompetenzen bei der Erstellung und Kuration von Daten erfor-

derlich, um deren Nachnutzung gewährleisten zu können. Diese Kompetenzen schließen u. a. eine genaue Kenntnis von fachlich einschlägigen Datenformaten ein, inklusive von Standards und Empfehlungen zur Datenmodellierung, die von Organisationen wie der International Standards Organisation (ISO), dem W3 Konsortium und im Fall von textbasierten Daten von der Text Encoding Initiative (TEI) verabschiedet worden sind. Diese Informationen sind besonders für Einzelforscher/innen und Nachwuchswissenschafler/-innen wichtig, die häufig (noch) nicht über die nötige Erfahrung in diesen Bereichen verfügen.

Die oben genannten Kompetenzen bei der Erstellung von digitalen Forschungsdaten, bei der Entwicklung von fachnaher Forschungssoftware und bei der zumeist fachnahen Entwicklung von Informationsangeboten für wissenschaftliche Daten haben nicht nur zu neuen Berufsfeldern von Datenmanager/-innen geführt. Sie gehen auch deutlich über die eher generischen, fachübergreifenden Informationsangebote von Bibliotheken und Rechenzentren hinaus. Daher sind in vielen Wissenschaftszweigen digitale Forschungsinfrastrukturen entstanden, um die notwendigen Informationsangebote zu entwickeln und nachhaltig anzubieten. Digitale Forschungsinfrastrukturen sind daher nicht nur ein auf die Fachdisziplin spezialisiertes Archiv für Forschungsdaten und lizensierte Daten, sondern als Forschungsinfrastruktur zugleich Ort und Werkzeug, um Forschung zu ermöglichen, kollaboratives Forschen zu unterstützen und fachnahe bzw. fachspezifische Forschungssoftware nachhaltig anzubieten. In Fachdisziplinen, in denen Forschungsdaten besonders vielfältig sind und daher eine große Bandbreite von Fachkompetenzen abdecken müssen wie im Fall der Sprachwissenschaft (vgl. Abschnitt 2), sind Forschungsinfrastrukturen nicht als zentrales Datenzentrum organisiert, sondern als ein Netzwerk von ortsverteilten Daten- und Kompetenzzentren, die jeweils unterschiedliche Spezialisierungen haben und in der Summe das Fach in hinreichender Breite abzudecken vermögen.

Der Wissenschaftsrat betont den Entwicklungsbedarf im Bereich der Infrastrukturen für die Geistes- und Sozialwissenschaften und ihre Rolle als „... von tradierenden und Fachinformation bevorratenden Hilfseinrichtungen zu Inkubatoren für neue und innovative wissenschaftliche Fragestellungen aufgrund von Forschungsdaten“ ebenso wie die Notwendigkeit, „... der Infrastrukturentwicklung für die Geistes- und Sozialwissenschaften in Deutschland mehr Aufmerksamkeit zu widmen“ (Wissenschaftsrat 2011: 8) um die internationale Anschlussfähigkeit der Forschung in den beteiligten Fachdisziplinen zu gewährleisten.

Der Aufbau von nachhaltigen Forschungsinfrastrukturen in allen Wissenschaftszweigen gehört zu den obersten Zielen der Europäischen Kommission seit dem 7. und 8. Rahmenprogramm und wird im aktuellen Horizon 2020

Rahmenprogramm mit hoher Priorität und mit flankierenden Maßnahmen wie der Einrichtung einer *European Open Science Cloud* weiter verfolgt. Die damit verbundenen Planungs- und Steuerungsprozesse hat die Europäische Kommission seit 2002 in die Hände des European Strategy Forum for Research Infrastructures (ESFRI) gelegt. Das ESFRI Board hat mit der ESFRI Roadmap 2006 erstmals einen Entwicklungsplan für den Aufbau von Forschungsinfrastrukturen vorgelegt, der alle Wissenschaftsbereiche umfasst und der kontinuierlich fortgeschrieben wird. In einem kompetitiven Auswahlverfahren wurden mit den Initiativen CLARIN (Hinrichs & Krauwer 2014) und DARIAH (Blanke et al. 2011) zwei Forschungsinfrastrukturen für die Geistes- und Sozialwissenschaften in die erste ESFRI Roadmap aus dem Jahr 2006 aufgenommen, die bis 2015 Gültigkeit hatte. Auf der aktualisierten ESFRI-Roadmap aus dem Jahr 2016 werden CLARIN und DARIAH als sog. *ESFRI Landmarks* geführt, d. h. als Forschungsinfrastrukturen, deren Aufbau schon weit fortgeschritten ist und die herausragende Serviceangebote für ihre Fachdisziplinen bereitstellen.

4 CLARIN: eine europäische Infrastruktur für die Geistes- und Sozialwissenschaften

4.1 ESFRI Roadmap und CLARIN ERIC

Die *Common Language Resources and Technology Infrastructure* (CLARIN) ist eine paneuropäische Forschungsinfrastruktur für die Geistes- und Sozialwissenschaften mit gegenwärtig über zwanzig Daten- und Kompetenzzentren in Europa und Nordamerika. Die CLARIN-Datenzentren stellen in der Summe ein umfangreiches, multilinguales Angebot an sprachbezogenen Forschungsdaten, digitalen Sprachdaten und digitalen Diensten für die akademische Forschung zur Verfügung.

Um die CLARIN-Forschungsinfrastruktur auf europäischer Ebene langfristig zu etablieren, hat CLARIN als eine der ersten ESFRI-Initiativen einen Antrag auf Einrichtung eines *European Research Infrastructure Consortium* (ERIC) gestellt. CLARIN bedient sich damit der juristischen Rechtsform, die von der Europäischen Kommission für den Betrieb von Forschungsinfrastrukturen eigens geschaffen wurde (COUNCIL REGULATION (EC) No 723/200922) und die inzwischen breite Akzeptanz seitens der nationalen Stakeholder gefunden hat. Das CLARIN ERIC wurde am 29. Februar 2012 durch eine Ratsentscheidung der Europäischen Kommission offiziell eingerichtet. Das CLARIN ERIC hat gegenwärtig (Stand: Mai 2017) neunzehn Mitglieder mit den Ländern Bulgarien,

Deutschland, Dänemark, Estland, Finnland, Griechenland, Italien, Lettland, Litauen, Niederlande, Norwegen, Österreich, Polen, Portugal, Schweden, Slowenien, Tschechien, Ungarn sowie die Niederländische Sprachunion als transnationaler Organisation. Als Beobachter sind Frankreich und Großbritannien am CLARIN ERIC beteiligt. Das Rückgrat der CLARIN-Infrastruktur besteht aus einem offenen Netzwerk von geographisch verteilten Datenzentren, die durch eine gemeinsame technische Infrastruktur verbunden sind und die sich regelmäßig dem externen Zertifizierungsprozess des *Data Seal of Approval* (Dillo & De Leeuw 2014) unterziehen.

4.2 Die nationale Forschungsinfrastruktur CLARIN-D

In Deutschland beteiligt sich das CLARIN-D Netzwerk (Hinrichs & Trippel 2017; www.clarin-d.eu) mit gegenwärtig zehn zertifizierten CLARIN-Zentren an der CLARIN-Forschungsinfrastruktur. Die nationale Koordination des CLARIN-D-Netzwerks erfolgt durch das CLARIN-D-Koordinationsbüro an der Universität Tübingen. Die CLARIN-D-Zentren verteilen sich auf neun verschiedenen Standorte in Deutschland: das Institut für Deutsche Sprache in Mannheim, das Zentrum Sprache der Berlin-Brandenburgische Akademie der Wissenschaften in Berlin sowie die Universitäten Duisburg-Essen, Hamburg, Leipzig, LMU München, Stuttgart und Tübingen. Ein weiteres CLARIN-D-Zentrum befindet sich am Max-Planck Institut (MPI) für Psycholinguistik in Nimwegen. Weitere wissenschaftliche Einrichtungen und Universitäten in Deutschland, darunter das LOEWE-Zentrum an der Universität Frankfurt, haben bereits Interesse geäußert, dem CLARIN-D-Zentrenverbund als zertifiziertes Datenzentrum beizutreten.

Der CLARIN-D-Zentrenverbund stellt sprachwissenschaftliche Forschungsdaten und digitale Sprachdaten in großem Umfang zur Verfügung. Es handelt sich dabei um gegenwartsbezogene und historische Korpusdaten geschriebener und gesprochener Sprache, multi-modale Daten, annotierte Daten, digitale Wörterbücher und Wortnetze, Sprachvariationsdaten, Archive bedrohter Sprachen, virtuelle Kollektionen von Sprachdaten sowie Forschungsprimärdaten von Einzelforschenden und Forschungsverbünden. Dazu gehören das in Abschnitt 2.2 erwähnte Deutsche Referenzkorpus (DeReKo), das Deutsche Textarchiv (DTA) sowie Korpora gesprochener Sprache des Bayerischen Archivs für Sprachsignale (BAS), des Instituts für Deutsche Sprache und des Hamburger Zentrums für Sprachkorpora (HZSK) als Sprachressourcen der CLARIN-D-Zentren in Hamburg, Mannheim und München.

Syntaktisch annotierte Textkorpora, sog. Baumbanken, wurden ursprünglich als Trainingsdaten für statistische Modelle in der Computerlinguistik entwickelt, werden inzwischen aber auch in der empirischen Grammatikforschung

und in der Korpuslinguistik genutzt. Für das Deutsche stellen die CLARIN-D-Zentren in Saarbrücken, Stuttgart und Tübingen mit der TIGER Baumbank und den Tübinger Baumbanken TüBa-D/Z, TüPP/D-Z und TüBa-D/S zum geschriebenen und gesprochenen Deutsch die derzeit meistgenutzten Baumbanken zum Gegenwartdeutschen zur Verfügung. Mit der Webapplikation TüNDRA (Martens 2013) bietet das CLARIN-D-Zentrum in Tübingen außerdem eine Recherche- und Visualisierungsplattform für Baumbanken an, in denen gegenwärtig Baumbanken für mehr als fünfzig verschiedenen Sprachen, darunter auch klassische Sprachen wie das Lateinische und Griechische, zugänglich sind.

Das CLARIN-D-Angebot an digitalen Wörterbüchern schließt das Projekt Deutscher Wortschatz (Quasthoff & Richter 2005) am CLARIN-D-Zentrum Leipzig, das deutsche Wortnetz GermaNet am CLARIN-D-Zentrum Tübingen und das Digitale Wörterbuch der deutschen Sprache (DWDS) am CLARIN-D-Zentrum an der BBAW Berlin ein.

Obwohl Forschungsdaten und digitale Sprachdaten zum Deutschen verständlicherweise eine besondere Rolle bei den Ressourcen der CLARIN-D-Zentren ausmachen, stellen die CLARIN-D-Zentren auch Sprachdaten zu weiteren Einzelsprachen und Sprachfamilien zur Verfügung, u. a. das DOBES Archiv für bedrohte Sprachen des CLARIN-D-Zentrums am MPI Nimwegen.

Die Forschungsinfrastruktur CLARIN-D möchte den gesamten Lebenszyklus von sprachbasierten Forschungsdaten und digitalen Sprachdaten unterstützen, indem es drei Kernangebote und damit verbundene Kompetenzen anbietet. Diese Kernangebote beziehen sich auf das Auffinden, Auswerten sowie Aufbereiten und Aufbewahren von Text- und Sprachdaten unterschiedlichster Datentypen und Provenienz.[4]

Unter Auffinden versteht CLARIN-D die Bereitstellung von Sprachressourcen mit dem Ziel, sie einer breiten Fachöffentlichkeit leicht zugänglich zu machen. Dies schließt die webbasierte Suche in standardkonformen Metadaten ebenso ein wie die Suche in den Sprachdaten und deren Annotationen. Zum Auffinden von Sprachressourcen stellt CLARIN das Virtual Language Observatory (VLO; van Uytvanck, Stehouwer & Lampen 2012) mit gegenwärtig über 900.000 Metadateneinträgen (Stand September 2016) bereit, in

4 Mit dem Datenlebenszyklus ist der zyklische Prozess beim Umgang mit Daten gemeint, der aus der Datengenerierung, der Datenaufbereitung, der Datenanalyse und der Datenarchivierung und -veröffentlichung besteht. Durch die Nachnutzung von Daten können neue Forschungsdaten entstehen, die die genannten Arbeitsschritte ebenfalls durchlaufen müssen und zu einem insgesamt zyklischen Prozess im Umgang mit Daten führen. Für eine hilfreiche Definition des Begriffs *Datenlebenszyklus* siehe Rat für Informationsinfrastrukturen (RfII) (2014: Anhang A).

dem die Nutzenden mit Hilfe einer Facettensuche gezielt nach digital verfügbaren Sprachdaten und -werkzeugen suchen können, die für ihr jeweiliges Forschungsvorhaben relevant sind.

Für die Suche in den von den CLARIN-D-Zentren bereit gestellten Sprachressourcen stehen spezialisierte, webbasierte Recherchewerkzeuge zur Verfügung, die mit Hilfe von Abfragesprachen die gezielte Suche nach Einzelwörtern, Kollokationen, Phrasen und/oder syntaktischen Konstruktionen ermöglichen. Webbasierte Suchfunktionen dieser Art stehen u. a. für das Deutsche Textarchiv, für die Sprachkorpora des IDS und für die Saarbrücker Textkorpora zur Verfügung. Es handelt sich dabei im Einzelnen um die Suchwerkzeuge DiaCollo (zur diachronen Kollokationsanalyse) und die linguistische Suchhilfe DDC für das Deutsche Textarchiv, die COSMAS-II$_{web}$ und KorAP Suchschnittstellen für das Deutsche Referenzkorpus und das CQPweb, das vom CLARIN-D-Zentrum Saarbrücken verwendet wird. Dadurch dass die CLARIN-Zentren durch eine gemeinsame technische Infrastruktur miteinander vernetzt sind, können die Sprachressourcen auch zentrenübergreifend nach relevanten Daten durchsucht werden. Dafür bietet CLARIN eine föderierte Inhaltssuche (engl.: *federated content search*) an, in denen Sammlungen von Sprachdaten, die in verschiedenen Repositorien archiviert sind, parallel und repositorienübergreifend abgefragt und die Suchergebnisse in einem gemeinsamen Ergebnis aggregiert werden können.

Unter Auswerten versteht CLARIN-D die Vielzahl an Werkzeugen, die den Forschenden zur Annotation und zur Analyse der Daten zur Verfügung gestellt werden. Hierbei können sowohl die von CLARIN-D angebotenen Sprachressourcen als auch eigene Daten der Nutzenden analysiert und durch eigene Annotationen angereichert werden. Für die manuelle Annotation bietet CLARIN-D Werkzeuge und virtuelle Forschungsumgebungen an, die auf spezifische Typen von Sprachdaten spezialisiert sind. Textuelle Sprachdaten können mit Hilfe der virtuellen Forschungsumgebung WebAnno (Yimam & Gurevych 2013) annotiert werden. Dabei können die zu bearbeitenden Daten bei Bedarf von mehreren Annotator/-innen genutzt und kollaborativ annotiert werden. Für multi-modale Sprachdaten stehen die Werkzeuge ELAN (Auer et al. 2010) und EXMERaLDA (Schmidt 2004) zur Verfügung, die es erlauben, Sprach- und Videoaufnahmen zu transkribieren, zu alignieren und zu annotieren.

CLARIN-D stellt neben Werkzeugen zur manuellen Annotation auch automatische Annotationswerkzeuge für Sprachdaten mit Hilfe von computerlinguistischen Verfahren zur Verfügung. Die dafür benötigten computerlinguistischen Werkzeuge, u. a. Tokenisierer, Lemmatisierer, morphologische Analyseprogramme, Wortartentagger, Parser und Eigennamenerkenner, werden u. a. von den CLARIN-D-Zentren in Berlin, München, Leipzig, Stuttgart

und Tübingen bereitgestellt. In der virtuellen Forschungsumgebung WebLicht (Hinrichs, Hinrichs & Zastrow 2010) werden diese Werkzeuge in einer gemeinsamen, webbasierten Benutzeroberfläche zusammengeführt. WebLicht erlaubt es, Sprachressourcen durch mehrschichtige Annotationsebenen anzureichern. Dazu werden die einzelnen computerlinguistischen Werkzeuge zu Verarbeitungsketten zusammengefügt und die Interoperabilität zwischen den Werkzeugen durch ein gemeinsames Textenkodierungsformat sichergestellt. Am Ende einer solchen Verarbeitungskette steht ein mit linguistischen Informationen angereichertes Textkorpus, das in Form eines XML-Dokuments weiter online analysiert, visualisiert oder heruntergeladen werden kann. WebLicht bietet eine Reihe vordefinierter Verarbeitungsketten an; Nutzende können mit Hilfe eines benutzerfreundlichen *drag-and-drop* Mechanismus aber auch eigene Verarbeitungsketten definieren und ausführen.

Für die automatische Alignierung von Sprachdaten und deren Transkriptionen hat das CLARIN-D-Zentrum in München die Webapplikation WebMAUS (Strunk, Schiel & Seifart 2014) entwickelt. Mit Hilfe der grafischen Benutzeroberfläche von WebMAUS können Forschende Dateien mit Sprachaufnahmen und die dazugehörigen Transkriptionen hochladen, die von WebMAUS automatisch aligniert werden. Die von WebMAUS alignierten Daten können dann bei Bedarf auf der Grundlage der Transkriptionen mit Hilfe von WebLicht um weitere Annotationsebenen angereichert werden.

Alle CLARIN-D-Annotationswerkzeuge müssen dem Forschungsstand entsprechend weiterentwickelt und gewartet werden, um die angebotenen Services an den aktuellen Stand der Wissenschaft anzupassen. Zusätzlich wird das bestehende Angebot in enger Absprache mit den von CLARIN-D adressierten Fachwissenschaften um weitere Softwarewerkzeuge kontinuierlich ergänzt.

Unter Aufbereiten und Aufbewahren versteht CLARIN-D das Verfügbarmachen von Forschungsdaten. Zur Datenaufbereitung stellt CLARIN-D dem Forschungsstand entsprechende Handreichungen, Expertisen und Konvertierungswerkzeuge zur Verfügung, die kontinuierlich aktualisiert und erweitert werden. Forschende können ihre Daten und zugehörige Metadaten in einem der CLARIN-D-Repositorien ablegen.

Um Forschende bei der Nutzung der CLARIN-Infrastrukturangebote zu unterstützen, stellen die CLARIN-D-Zentren verschiedene Beratungsdienste zur Verfügung. Das CLARIN-D-Zentrum in Hamburg koordiniert eine Beratungsplattform (Lehmberg 2014), die individuelle Anfragen zu CLARIN-D-Ressourcen, virtuellen Forschungsumgebungen, Werkzeugen und Diensten mit Hilfe eines Ticketsystems an Expert/-innen in den fachlich einschlägigen CLARIN-D weiterleitet. Im Netz verfügbare Screencasts, Online-Tutorien und FAQ-Listen, die in die CLARIN-D-Beratungsplattformen integriert sind, stellen über den

individuellen Nutzersupport hinausgehende Informationsangebote dar. Das CLARIN-D-Zentrum am IDS Mannheim bietet mit dem Legal Helpdesk (Ketzan & Kamocki 2012) allgemeine Informationen in Bezug auf ethische und juristische Aspekte bei der Erstellung und Nutzung von Forschungsdaten und digitalen Sprachressourcen an.

Als wertvolle Informationsquelle zu best-practise Richtlinien und für Sprachressourcen relevante Enkodierungsstandards haben Forschende Zugang zum CLARIN-D User Guide (Herold & Lemnitzer 2012), einem praktischen Leitfaden für die Erstellung neuer digitaler Sprachdaten und für die Anpassung und Integration existierender Sprachressourcen in die CLARIN-D-Infrastruktur.

Der Leitfaden adressiert u. a. die Empfehlungen der Deutschen Forschungsgemeinschaft zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora, an deren Erstellung Expert/-innen aus den CLARIN-D-Zentren beteiligt waren.

4.3 Nutzungszahlen und beteiligte Fachdisziplinen

Da die CLARIN Forschungsinfrastruktur Sprachressourcen und -dienste anbietet, bilden Forschende aus dem Bereich der Sprachwissenschaft einen natürlichen Nutzerkreis. Andererseits sind sprachbezogene Ressourcen und Dienste auch für weitere geistes- und sozialwissenschaftliche Fachdisziplinen von hoher Relevanz, für die sprachliche Daten eine wichtige empirische Grundlage darstellen. Dies trifft in besonderer Weise für Forschende im Bereich der Digital Humanities zu, einer Forschungsrichtung, die digitale Ressourcen und datenbasierte Methoden verwendet, um neue Antworten auf zentrale Fragen ihres Fachs zu geben oder gänzlich neue Forschungsfragen zu bearbeiten. Die konstant große Resonanz bei einschlägigen nationalen und internationalen Fachkonferenzen und die stark ansteigende Anzahl von einschlägigen Studiengängen und Studienangeboten unterstreicht die wissenschaftliche Produktivität und die Attraktivität der Digital Humanities. Um die CLARIN-D Forschungsinfrastruktur an den Bedarfen der adressierten Fachdisziplinen auszurichten, kooperieren die CLARIN-D Zentren mit Facharbeitsgruppen, denen gegenwärtig mehr als zweihundert Fachwissenschaftler/-innen aus der Sprach-, Kognitions-, Geschichts- und Politikwissenschaft angehören. Im Bereich der Sprachwissenschaft engagieren sich Forschende aus der Allgemeinen Sprachwissenschaft, der Quantitativen Linguistik, der Germanistik, der Anglistik, der Romanistik, der Slawistik, der Phonetik und Phonologie, der Computerlinguistik und der Psycholinguistik in den CLARIN-D Facharbeitsgruppen.

Das CLARIN-D Angebot wird von zahlreichen Einzelforschenden und Forschungsverbünden genutzt. Dies ist unter anderem an den Kooperationspro-

jekten abzulesen, die CLARIN-D Zentren mit anderen Institutionen unterhalten. Zusammengenommen kooperieren die zehn CLARIN-D Zentren gegenwärtig mit 462 externen Forschungsprojekten, die sich auf insgesamt 35 Fachdisziplinen erstrecken. Der Schwerpunkt bei den beteiligten Fachdisziplinen liegt dabei in den Geistes- und Sozialwissenschaften, erstreckt jedoch auch auf die Verhaltens- und Naturwissenschaften, speziell auf die Neurowissenschaften und die Informatik. Die CLARIN-D Kooperationspartner beschränken sich dabei nicht nur auf Universitäten, Hochschulen und außeruniversitäre Forschungseinrichtungen in Deutschland und im deutschsprachigen Raum. Die deutschsprachigen Ressourcen von CLARIN-D spielen weltweit für die Auslandsgermanistik eine wichtige Rolle. Besonders erwähnenswert ist dabei die Tatsache, dass sich die Nutzung der CLARIN-D Infrastruktur nicht nur auf die Forschung beschränkt. So werden die Annotationswerkzeuge, die durch die virtuelle Forschungsumgebung WebLicht bereitgestellt werden, auch für die universitäre Lehre eingesetzt, um das Lehrangebot im Bereich Grammatik durch die digitalen Angebote zu erweitern.

Um die Nutzung von CLARIN-D Ressourcen für die Geistes- und Sozialwissenschaften abschätzen zu können, sammelt CLARIN-D Daten über die Anzahl der registrierten Wissenschaftler/-innen von CLARIN-D-Sprachressourcen und über die Nutzung von webbasierten CLARIN-D-Diensten. Bisher wurden die CLARIN-D-Sprachdaten von mehr als 33.000 Wissenschaftler/-innen aus 2.421 akademischen Institutionen aus dem In- und Ausland genutzt (Stand: April 2017). Für die webbasierten CLARIN-D-Dienste wurden im Zeitraum von 2014 bis April 2017 insgesamt 3.190.949 Aufrufe gezählt. Diese Nutzerstatistiken zeigen, dass die CLARIN-D-Forschungsinfrastruktur in vielfältiger Weise von einer großen Anzahl von Forschenden intensiv genutzt wird.

5 Schlussbemerkungen und Ausblick

Ich komme abschließend zum Ausgangspunkt meiner Überlegungen zurück. Die Vielfalt und Bedeutung empirischer und experimenteller Forschungsdaten unterstreicht die Relevanz von Forschungsinfrastrukturen für die Sprachwissenschaft. Aus der Sicht anderer geisteswissenschaftlicher Disziplinen wird auf die Linguistik häufig als besonders gutes Beispiel für den Umgang mit wissenschaftlichen Daten verwiesen. Auf diesem Hintergrund werden Infrastrukturangebote für die Sprachwissenschaft eine wichtige und in der Zukunft noch wichtigere Säule für das Fach bilden.

Literatur

Auer, Eric, Albert Russel, Han Sloetjes, Peter Wittenburg, Oliver Schreer, S. Masnieri, Daniel Schneider & Sebastian Tschöpel (2010): ELAN as flexible annotation framework for sound and image processing detectors. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (Hrsg.), *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta: European Language Resources Association (ELRA).

Baker, Collin F., Charles J. Fillmore & John B. Lowe (1998): The Berkeley Framenet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, 86–90. Montreal, Quebec, Canada: Association for Computational Linguistics. doi: 10.3115/980845.980860. http://www.aclweb.org/anthology/P98-1013.

Blanke, Tobias, Michael Bryant, Mark Hedges, Andreas Aschenbrenner & Michael Priddy (2011): Preparing DARIAH. In *E-Science (e-Science), 2011 IEEE 7th International Conference*, 158–165. DOI: 10.1109/eScience.2011.65.

Bloomfield, Leonhard (1914): *Introduction to the Study of Language*. Henry Holt.

Brants, Thorsten (2000): TnT – A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, 224–231. Seattle, Washington, USA: Association for Computational Linguistics. doi: 10.3115/974147.974178. http://www.aclweb.org/anthology/A00-1031.

Bresnan, Joan (2016): Linguistics: The Garden and the Bush: Write-up of ACL Lifetime Achievement Award Acceptance Speech for 2016. *Computational Linguistics* 42(4), 599–617.

Brown, Peter F., John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer & Paul S. Roossin (1990): A Statistical Approach to Machine Translation. *Computational Linguistics* 16(2), 79–85.

Burnard, Lou & Guy Aston (1998): *The BNC handbook: Exploring the British National Corpus*. Edinburgh: Edinburgh University Press.

Chomsky, Noam (1957): *Syntactic Structures*. The Hague: Mouton.

Church, Kenneth Ward (1988): A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, 136–143. Austin, Texas, USA: Association for Computational Linguistics.

Collins, Michael (2003): Head-driven Statistical Models for Natural Language Parsing. *Computational Linguistics* 29(4), 589–637.

Dillo, I. & Lisa De Leeuw (2014): *Data Seal of Approval: Certification for sustainable and trusted data repositories*. Data Archiving and Networked Services (DANS).

Eichinger, Ludwig (1982): *Syntaktische Transposition und semantische Derivation. Die Adjektive auf -isch im heutigen Deutsch.* Linguistische Arbeiten 113. Tübingen: Niemeyer.

Erk, Katrin, Andrea Kowalski, Sebastian Padó & Manfred Pinkal (2003): Towards a Resource for Lexical Semantics: A large German corpus with extensive semantic annotation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 537–544. Sapporo, Japan: Association for Computational Linguistics. doi: 10.3115/1075096.1075164. http://www.aclweb.org/anthology/P03-1068.

Geyken, Alexander, Jörg Didakowski & Alexander Siebert (2008): Generation of Word Profiles on the Basis of a Large and Balanced German Corpus. In *Proceedings of the XIII EURALEX International Congress. Euralex 2008*, 371–385.

Geyken, Alexander, Susanne Haaf, Bryan Jurish, Matthias Schulz, Jakob Steinmann, Christian Thomas & Christian Wiegand (2011): Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv. In Silke Schomburg, Claus Leggewie, Henning Lobin & Cornelius Puschmann (Hrsg.), *Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland. Beiträge der Tagung*, 157–161. Köln: Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen (hbz).

Gray, Russell D., Alexei J. Drummond & Simon J. Greenhill (2009): Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement. *Science* 323(5913), 479–483. doi: 10.1126/science.1166858.

Henrich, Verena & Erhard Hinrichs (2010): GernEdiT – the GermaNet Editing Tool. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (Hrsg.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta: European Language Resources Association (ELRA).

Herold, Axel & Lothar Lemnitzer (2012): CLARIN-D User Guide. Tech. rep. Berlin-Brandenburgische Akademie der Wissenschaften, Zentrum Sprache Berlin. Online verfügbar: http://media.dwds.de/clarin/userguide/userguide-1.0.1.pdf.

Hinrichs, Erhard (2016): Substitute Infinitives and Oberfeld Placement of Auxiliaries in German Subordinate Clauses: A synchronic and diachronic corpus study using the CLARIN research infrastructure. *Lingua* 178, 46–70.

Hinrichs, Erhard (2017): Morphological productivity of adjective formation in German – a diachronic corpus study using the CLARIN-D infrastructure. In *The Book of Abstracts of the CLARIN 2017 Annual Conference*, Budapest. https://www.clarin.eu/content/abstracts-overview-clarin-annual-conference-2017.

Hinrichs, Erhard, Marie Hinrichs & Thomas Zastrow (2010): WebLicht: Web-Based LRT Services for German. In *Proceedings of the ACL 2010 System Demonstrations*, 25–29. Uppsala, Sweden: Association for Computational Linguistics.

Hinrichs, Erhard & Steven Krauwer (2014): The CLARIN Research Infrastructure: Resources and Tools for eHumanities Scholars. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (Hrsg.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 1525–1531. Reykjavik, Iceland: European Language Resources Association (ELRA).

Hinrichs, Erhard & Thorsten Trippel (2017): Clarin-D: eine Forschungsinfrastruktur für die sprachbasierte Forschung in den Geistes- und Sozialwissenschaften. *Bibliothek Forschung und Praxis* 41(1), 45–54.

Jelinek, Frederick (1997): *Statistical Methods for Speech Recognition*. Cambridge, MA: MIT Press.

Ketzan, Erik & Pawel Kamocki (2012): The CLARIN-D Legal Help Desk and Emerging Copyright Issues for Language Scientists. In *Proceedings of LREC 2012 Workshop on Legal Issues*.

Kilgarriff, Adam, Pavel Rychlý, Pavel Smrž & David Tugwell (2004): The Sketch Engine. In *Proceedings of the 11th EURALEX International Congress*, 105–116. Lorient, France.

Klavans, Judith L. & Phil Resnik (Hrsg.) (1997): *The Balancing Act: Combining symbolic and statistical approaches to language*. Cambridge, MA: The MIT Press.

Klein, Wolfgang & Alexander Geyken (2010): Das Digitale Wörterbuch der Deutschen Sprache (DWDS). In *Lexikographica*, 79–93. Berlin/New York: De Gruyter.
Kupietz, Marc, Cyril Belica, Holger Keibel & Andreas Witt (2010): The German Reference Corpus DeReKo: A primordial sample for linguistic research. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (Hrsg.), *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta: European Language Resources Association (ELRA).
Lehmberg, Timm (2014): The CLARIN-D Helpdesk. In *Papers, Posters and Demos CAC2014*, CLARIN ERIC: Utrecht, The Netherlands. Available at: https://www.clarin.eu/sites/default/files/cac2014_submission_23_0.pdf.
Martens, Scott (2013): TüNDRA: A Web Application for Treebank Search and Visualization. In Sandra Kübler, Petya Osenova & Martin Volk (Hrsg.), *Proceedings of the Twelfth Workshop on Treebanks and Linguistic Theories (TLT 12)*, 133–144. Sofia: Institute of Information and Communication Technologies. Bulgarian Academy of Sciences.
Meibauer, Jörg & Markus Steinbach (Hrsg.) (2011): *Experimental Pragmatics/Semantics*. Amsterdam: John Benjamins Pub. Co.
Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross & Katherine Miller (1990): Wordnet: An on-line lexical database. *International Journal of Lexicography* 3, 235–244.
Noveck, Ira A. & Dan Sperber (Hrsg.) (2004): *Experimental Pragmatics*. Palgrave Studies in Pragmatics, Language and Cognition. Basingstoke, Hampshire: Palgrave Macmillan.
Pierrehumbert, Janet, Mary E. Beckman & D. Robert Ladd (2000): Conceptual Foundations in Phonology as a Laboratory Science. In Noel Burton-Roberts, Philip Carr & Gerard Doherty (Hrsg.), *Phonological Knowledge: Conceptual and Empirical Issues*, 273–303. Oxford: Oxford University Press.
Quasthoff, Uwe & Matthias Richter (2005): Projekt Deutscher Wortschatz. *Babylonia* 03, 33–35.
Rat für Informationsinfrastrukturen (RfII) (2014): Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland. Tech. rep. http://www.rfii.de/download/rfii-empfehlungen-2016/.
Sapir, Edward (1921): *Language: An Introduction to the Study of Speech*. Harcourt, Brace.
de Saussure, Ferdinand (1916): *Cours de linguistique générale*. Payot.
Schmid, Helmut (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.
Schmidt, Thomas (2004): Exmaralda – ein Modellierungs- und Visualisierungsverfahren für die computergestützte Transkription gesprochener Sprache. In Ernst Buchberger (Hrsg.), *Proceedings of Konvens 2004*, Wien: Schriftenreihe der Österreichischen Gesellschaft für Artificial Intelligence.
Schütze, Carston (1996): *The Empirical Base of Linguistics: Grammaticality judgments and linguistic methodology*. Chicago, Illinois: Chicago University Press.
Sprouse, Jan & Norbert Hornstein (Hrsg.) (2012): *Experimental Syntax and Island Effects*. Cambridge University Press.
Strunk, Jan, Florian Schiel & Frank Seifart (2014): Untrained Forced Alignment of Transcriptions and Audio for Language Documentation Corpora using WebMAUS. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (Hrsg.), *Proceedings of*

the Ninth International Conference on Language Resources and Evaluation (LREC'14), 3940–3947. Reykjavik, Iceland: European Language Resources Association (ELRA).

van Uytvanck, Dieter, Herman Stehouwer & Lari Lampen (2012): Semantic Metadata Mapping in Practice: the Virtual Language Observatory. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk & Stelios Piperidis (Hrsg.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, 1029–1034. Istanbul, Turkey: European Language Resources Association (ELRA).

Wissenschaftsrat (2011): *Empfehlungen zu Forschungsinfrastrukturen in den Geistes- und Sozialwissenschaften* Drs. 10465-11. Berlin: Presse- und Öffentlichkeitsarbeit des Wissenschaftsrates.

Yimam, Seid Muhie & Iryna Gurevych (2013): WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. In *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics (ACL) – System Demonstrations*, 1–6.

Stefan Schmunk, Frank Fischer, Mirjam Blümm und
Wolfram Horstmann

3 Interoperabel und partizipativ

Digitale Forschungsinfrastrukturen in den Geisteswissenschaften am Beispiel von DARIAH-DE und DARIAH-EU

Abstract: DARIAH baut seit 2006 in Europa und seit 2011 in Deutschland eine digitale Forschungsinfrastruktur für die Geistes- und Kulturwissenschaften auf. Im Zentrum der Entwicklung der kommenden Jahre steht im nationalen Kontext die Überführung des Projekts DARIAH-DE hin zu einer Organisation und auf europäischer Ebene die Verstetigung des im Jahre 2014 gegründeten DARIAH-ERIC als *architecture of participation* unter Beteiligung der Fachcommunitys. Der Aufsatz beleuchtet die Entwicklungen von DARIAH und untersucht die historisch gesehen kurze Geschichte des Instruments Forschungsinfrastrukturen – eine wissenschaftliche Innovation des 20. Jahrhunderts – in den Geistes- und Kulturwissenschaften.

Keywords: Forschungswerkzeuge, Interoperabilität, Kooperatives Modell, Vernetzung

1 Einleitung

Die Digitalisierung der Gesellschaft ist ein anhaltender Prozess, der auch in den Geisteswissenschaften voranschreitet. Wenn im Folgenden dezidiert von einer digitalen Forschungsinfrastruktur für die Geistes-, Kultur- und Sozialwissenschaften (die *Humanities*) die Rede ist, dann bedeutet dies nicht, dass die Forschung sich ohne eine solche Infrastruktur nicht bereits umfassend digita-

Stefan Schmunk, University of Applied Sciences Darmstadt, Max-Planck-Str. 2,
D-64807 Dieburg, E-Mail: stefan.schmunk@h-da.de
Frank Fischer, National Research University Higher School of Economics, School of Linguistics,
Ul. Staraya Basmannaya 21/4, of. 207, RU-105066 Moskau, E-Mail: ffischer@hse.ru
Mirjam Blümm, Niedersächsische Staats- und Universitätsbibliothek Göttingen,
Platz der Göttinger Sieben 1, D-37073 Göttingen, E-Mail: bluemm@sub.uni-goettingen.de
Wolfram Horstmann, Niedersächsische Staats- und Universitätsbibliothek Göttingen,
Platz der Göttinger Sieben 1, D-37073 Göttingen, E-Mail: horstmann@sub.uni-goettingen.de

https://doi.org/10.1515/9783110538663-004

lisiert hätte. Auch wenn eine Wissenschaftlerin oder ein Wissenschaftler nicht dezidiert ‚digital forscht', nicht bewusst virtuelle Forschungsumgebungen und keine Software jenseits der Textverarbeitung benutzt, ist eine ‚undigitale' Forschung kaum mehr denkbar.

Die Einbeziehung digitaler Methodiken in den Alltag hat für die Geisteswissenschaften eine interessante Folge. Jenseits hochspezialisierter Forschungsfragen in den jeweiligen Einzeldisziplinen kann fachübergreifend gemeinsam über digitale Arbeitsweisen diskutiert und können vielseitige Lösungen entwickelt werden. Zur traditionellen inhaltlichen Interdisziplinarität der Geistes- und Kulturwissenschaften gesellt sich eine praktische. Da durch die Inkorporation fachfremder Epistemologien, etwa der Informatik oder Mathematik, der Komplexitätsgrad steigt, ändert sich auch der Phänotyp der Forschung. Die Arbeit im Team wird zum Standard, und statt des Einzelaufsatzes oder der im Alleingang verantworteten Monografie organisiert sich Wissensgewinn in Projekten und durch Kollaboration. Diese neuen Aspekte der Zusammenarbeit bilden sich auch im Aufbau digitaler Forschungsinfrastrukturen ab, und hier kommt DARIAH ins Spiel, die *DigitAl Research Infrastructure for the Arts and Humanities*.

DARIAH versteht sich als technische und soziale Infrastruktur, als ein „social marketplace for services" (Blanke et al. 2011: 158 f.). Ihre Aufgabe besteht darin, Forscherinnen und Forscher bei ihrer Arbeit zu unterstützen, indem sie verlässliche Strukturen für die digitalen Aspekte des Forschungsalltags bereitstellt. Dabei kann es sich um Softwarelösungen handeln, etwa Werkzeuge wie TextGrid[1] (eine Allround-Suite für digitale Editionen) oder den Geo-Browser[2] (einen Online-Dienst für das Beforschen und Publizieren temporal-spatialer Daten).[3] DARIAH ist aber auch eine soziale und didaktische Plattform, über die Kenntnisse der digitalen Forschungspraxis vermittelt werden.

2 Digitale Forschungsinfrastrukturen

> Eine **Infrastruktur** (von lateinisch *infra* 'unterhalb' und *structura* 'Zusammenfügung') ist im übertragenen Sinn ein Unterbau. Sie umfasst alle langlebigen Einrichtungen materieller oder institutioneller Art, die das Funktionieren einer arbeitsteiligen Volkswirtschaft begünstigen. (Wikipedia 2017)

1 https://textgrid.de (letzter Zugriff: 20. 10. 2017).
2 https://geobrowser.de.dariah.eu (letzter Zugriff: 20. 10. 2017).
3 Eine Übersicht zu den Infrastrukturangeboten von DARIAH-DE und TextGrid findet sich in: Blümm, Funk & Söring (2015).

Der Begriff Infrastruktur durchlebte in den letzten Jahrzehnten eine enorme Konjunktur, die dazu führte, dass der Terminus mittlerweile nicht nur in technischen und technologischen Komplexen, sondern nahezu in allen denkbaren sozio-ökonomischen und politischen Bereichen Einzug gehalten hat. Er zählt spätestens seit den 1950er Jahren nicht nur zu einem festen sprachübergreifenden Vokabular, sondern ist vielmehr als einer der zentralen Begriffe der Moderne, die sich im Kontext der Sozialstaatlichkeit und des westeuropäischen und US-amerikanischen Wirtschaftswachstums in den Nachkriegsjahren des 20. Jahrhunderts durchsetzte, zu sehen. Der Bau von Straßen und Eisenbahnstrecken, deren wachsende Vernetzung und die damit verbundene Erschließung des Raumes, kurzum die sozioökonomische Modernisierung in den Jahren der Kriegs- und Nachkriegszeit prägten dieses neue technologisch-ökonomische Phänomen und damit auch den Begriff. Eine besondere Rolle spielten hierbei die Erschließung durch Elektrizität, mit Telegraphen- und später Telefonleitungen, fließendem Wasser und der Bau von Abwasserkanälen – dies im Gegensatz zum 19. Jahrhundert nicht mehr ausschließlich auf einen städtischen Raum beschränkt, sondern vielmehr als Erschließung des gesamten Landes, bis in den letzten Weiler und Kirchturm.[4] Die gesellschaftliche und vor allem die ökonomische Modernisierung und, damit einhergehend, das Wirtschaftswachstum im Zeitalter der Moderne, waren hierbei die Triebkräfte, die auch nach neuen Begrifflichkeiten und Termini verlangten. Aus diesem Grund ist es auch nicht verwunderlich, dass der Begriff „Infrastruktur“ etymologisch gegen Mitte des 19. Jahrhunderts in wirtschaftshistorischen Kontexten erstmals verwendet wurde, aber der unaufhaltsame Durchbruch erst Mitte des 20. Jahrhunderts einsetzte.

Abbildung 3.1 zeigt ein Diagramm mit der Häufigkeitskurve für den Begriff „Infrastruktur“, erstellt mit dem Ngram-Viewer von Google. Hierbei ist erkennbar, dass ab Mitte der 1940er Jahre der Begriff „Infrastruktur“ verwendet wird und die Nutzung in der Zeit des westdeutschen Wirtschaftswachstums, dem Erhard’schen „Wirtschaftswunder“ der 1950er Jahre, rapide ansteigt. Vergleichbare statistische Tendenzen und Verläufe zeigen sich auch im englischsprachigen Raum für den Terminus „infrastructure“.[5] Auch im französischen Sprachgebrauch steigt die Verwendungshäufigkeit ab Mitte der 1950er Jahre an, sodass von einem sprachübergreifenden Phänomen gesprochen werden kann

---

4 Eine ausgezeichnete Beschreibung dieser grundsätzlichen sozialen und gesellschaftlichen Veränderungen, die alle westeuropäischen Länder und die USA betrafen, findet sich bei Mak (1999).

5 Siehe: https://books.google.com/ngrams/graph?content=Infrastructure&year_start=1800&year_end=2008&corpus=18&smoothing=1 (letzter Zugriff: 20.10. 2017).

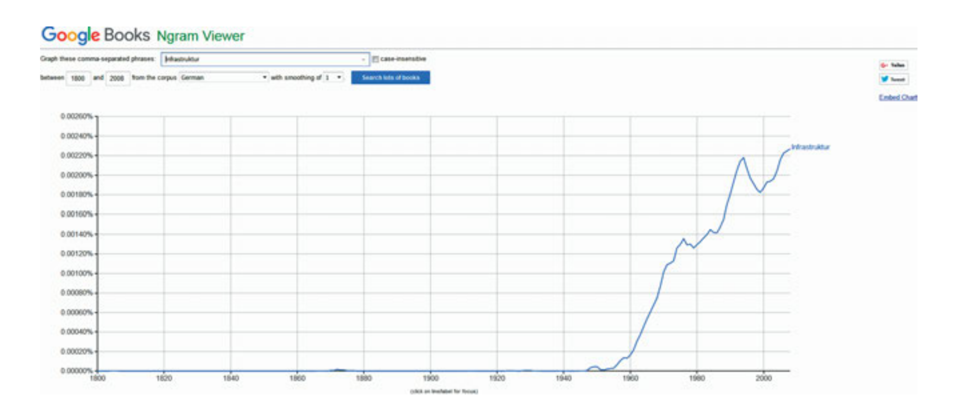


Abb. 3.1: Google-Ngram-Viewer: Visualisierung der relativen Häufigkeit des Begriffs „Infrastrukturen“ im deutschsprachigen Buchkorpus bei Google.[6]

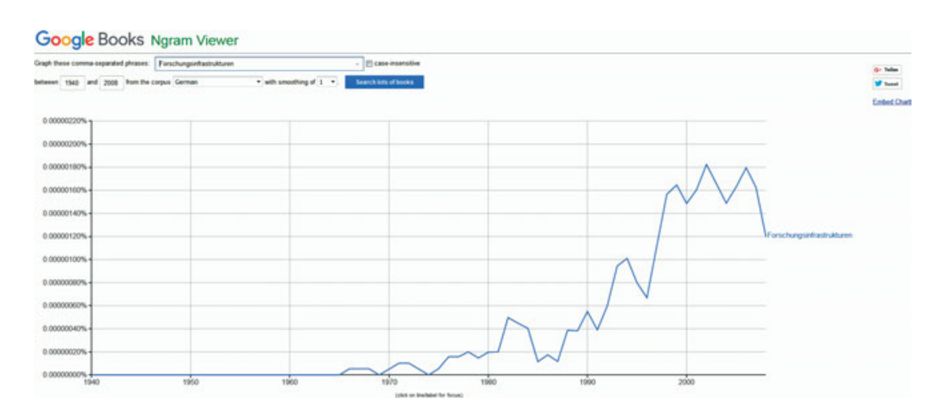


Abb. 3.2: Google-Ngram-Viewer: Häufigkeitskurve für den Begriff „Forschungsinfrastrukturen“ im deutschsprachigen Buchkorpus bei Google.[7]

und, statistisch gesehen, vergleichbare und parallele Entwicklungen im Deutschen, Englischen und auch Französischen erkennbar sind.

Von „Forschungsinfrastrukturen“ ist im Deutschen zu diesem Zeitpunkt übrigens noch keine Rede, zumindest lassen sich hierfür keine Nennungen finden, sondern, mit einer statistischen Variabilität, erst ab Ende der 1970er Jahre.

6 Siehe: https://books.google.com/ngrams/graph?content=Infrastruktur&year_start=1800&year_end=2008&corpus=20&smoothing=1 (letzter Zugriff: 20. 10. 2017).
7 Siehe https://books.google.com/ngrams/graph?content=Forschungsinfrastrukturen&year_start=1940&year_end=2008&corpus=20&smoothing=1 (letzter Zugriff: 20. 10. 2017).

Ganz gegensätzlich sind hier die Nennungen im Englischen, die bereits Ende der 1960er und insbesondere ab den 1970er Jahren einen rasanten Anstieg dieses Begriffes verzeichnen.[8] Dies ist auf die politischen Diskussionen der Europäischen Kommission zurückzuführen, die bereits 1972 die ersten Leitlinien für eine Gemeinschaftspolitik im Bereich von Forschung, Entwicklung und Innovation vorschlug, was letztlich etwa zehn Jahre später, 1983, zur Verabschiedung des ersten gemeinschaftlichen Rahmenprogramms führte (vgl. Europäisches Parlament 2016: 1 f., 6 f.). Die gemeinsame, auf Ebene der Europäischen Gemeinschaft initiierte Forschungspolitik hatte zwei Grunddimensionen zum Ziel: Einerseits sollte auf europäischer Ebene eine Koordinierung der einzelstaatlichen Forschungsstrategien erfolgen, andererseits der Fokus auf eine Verbesserung der Zusammenarbeit der Mitgliedstaaten im Bereich von Forschung und Entwicklung gelegt werden (vgl. Europäisches Parlament 2016: 6 f.). Als zu Beginn der 1980er Jahre das erste Rahmenprogramm geplant und strukturiert wurde, lag der Fokus vor allem auf der wirtschaftlichen Stärkung Europas durch Forschungs- und Entwicklungsvorhaben und damit verbunden auf der Initiierung von Infrastruktur- und Forschungsinfrastrukturvorhaben, die mitgliedstaatsübergreifend initiiert, entwickelt und betrieben werden sollten (vgl. Europäisches Parlament 2016: 8–10). Ein Schwerpunkt lag hierbei auf den angewandten Wissenschaften und den Naturwissenschaften, die bis zum heutigen Zeitpunkt oftmals den Begriff Forschungsinfrastrukturen konnotieren. Darin erzielte Forschungsergebnisse sollten mittels eines Wissenstransfers mittel- und langfristig in kommerzielle Bereiche transferiert und durch Industrie und Gewerbe genutzt werden können. Insbesondere sind Forschungsinfrastrukturen mit Großforschungsvorhaben sprachlich konnotiert, wie beispielsweise dem CERN (*Conseil Européen pour la Recherche Nucléaire*), das bereits 1953 gegründet wurde und in dem physikalische Grundlagenforschung betrieben wird. Eine Reihe weiterer naturwissenschaftlicher Großforschungsvorhaben sind auf europäischer Ebene im Rahmen der ESFRI-Prozesse (*European Strategy Forum on Research Infrastructures*) zu nennen, die in den letzten Jahren im Rahmen der Forschungsrahmenprogramme entwickelt und aufgebaut wurden, von denen derzeit fünf in den Geistes-, Kultur- und Sozialwissenschaften (SSH) zu verorten sind.[9] Bei diesen handelt es sich um SHARE[10], European

---

8 Siehe https://books.google.com/ngrams/graph?content=research+infrastructures&year_start=1950&year_end=2000&corpus=15&smoothing=3 (letzter Zugriff: 20. 10. 2017).

9 Eine genaue Aufstellung findet sich unter: http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri (letzter Zugriff: 20. 10. 2017).

10 The Survey of Health, Ageing and Retirement in Europe (SHARE): http://www.share-project.org/home0.html (letzter Zugriff: 20. 10. 2017).

Social Survey[11], CESSDA[12], CLARIN[13] und DARIAH[14]. Während in den Anfangsjahren der europäischen Forschungsförderung der Fokus vor allem auf naturwissenschaftlichen Forschungsvorhaben, der Grundlagenforschung und kommerziell nutzbaren Themen lag, werden seit ungefähr 15 Jahren verstärkt auch geistes- und sozialwissenschaftliche Vorhaben gefördert.

Forschungsinfrastrukturen sind, historisch gesehen – und wie aufgezeigt –, ein Novum des 20. Jahrhunderts und aufgrund ihrer Entwicklungsgeschichte vor allem naturwissenschaftlich oder klassisch infrastrukturell (Straßen, Verkehr, Energie, Telekommunikation etc.) ausgerichtet. In Deutschland selbst wurde dieses Themenfeld vor allem durch den Wissenschaftsrat ab 2007 aufgegriffen und vier Jahre später als *Empfehlungen zu Forschungsinfrastrukturen in den Geistes- und Sozialwissenschaften* thematisiert und publiziert (Wissenschaftsrat 2011). Dies stellte einen Paradigmenwechsel dar – und damit verbunden auch die Initiierung eines forschungskulturellen Wandels in den Geistes- und Kulturwissenschaften –, da zuvor der Fokus vor allem auf naturwissenschaftlichen Großgeräten und auf Fachdisziplinen lag, deren Forschungen von umfangreichen und kostenintensiven Apparaten und Einrichtungen abhängig waren. Das Ziel war hierbei, „geeignete Infrastrukturen und Schritte zu ihrer Entwicklung und Förderung [zu] identifizieren, die den Geistes- und Sozialwissenschaften in Deutschland optimale Bedingungen für international beachtenswerte Forschungen bieten [sollten]“ (Wissenschaftsrat 2011: 5 f.).

Die Herausforderung bestand und besteht darin, dass ein forschungspolitisches Konzept auf geistes-, kultur- und sozialwissenschaftliche Fachdisziplinen übertragen werden sollte, die sich eigentlich in ihren eigenen Fachtraditionen dadurch auszeichneten, dass die überwiegende Mehrzahl der Fachvertreterinnen und Fachvertreter eben keine Großgeräte für ihre Forschungen benötigten und die wenigsten von ihnen kollaborativ mit anderen forschten und publizierten. Es war ein „digitaler Einbruch“ in eine „Welt der klassischen Gelehrten“, in der bis zum heutigen Tag Forschungsleistungen zumeist immer noch als individuelle wissenschaftliche „Leistung“ betrachtet werden, die wiederum als Publikation in den wissenschaftlichen Diskurs der Fachdisziplin zurückgeführt wird. In vielerlei Hinsicht war die Nutzung von Forschungsinfrastrukturen zumeist auf die Nutzung von Bibliotheken, Archiven, Galerien und Museen und die Nutzung von Computern als digitale Schreib-

11 http://www.europeansocialsurvey.org (letzter Zugriff: 20. 10. 2017).
12 Consortium of European Social Science Data Archives (CESSDA): https://www.cessda.eu (letzter Zugriff: 20. 10. 2017).
13 https://www.clarin.eu (letzter Zugriff: 20. 10. 2017).
14 http://dariah.eu (letzter Zugriff: 20. 10. 2017).

geräte beschränkt. Der diesem Umstand zugrunde liegende Forschungskreislauf, der sich als Methodeninstrumentarium über lange Jahrzehnte entwickelte – wenn nicht sogar eine jahrhundertealte methodologische Evolution darstellt –, war zumeist klar definiert und der in den einzelnen Fächern entwickelte Kanon an Methoden und Theorien zumeist auf die Analyse des geschriebenen Wortes in Form von Büchern oder Archivalien abgestimmt. Kurzum, es war ein klassischer Forschungskreislauf der Hermeneutik, also dem Lesen, Verstehen, Analysieren, Interpretieren von Texten und der Generierung von Forschungsergebnissen in Form von neuen Texten.

Und dennoch sind bereits seit Mitte der 1970er Jahre erste Forschungsvorhaben sowie Wissenschaftlerinnen und Wissenschaftler zu nennen, die sich explizit mit digitalen Methoden bzw. mit digitalen Daten auseinandergesetzt haben, wie beispielsweise Manfred Thaller und das datenbankorientierte Programmiersystem CLIO (vgl. Grotum 2004: 37 f.) oder das Projekt *TUSTEP* (das *TUebinger System von TExtverarbeitungs-Programmen*)[15], das seit 1978 die elektronische Auszeichnung und Erschließung von Texten ermöglicht, um beispielsweise textkritische Ausgaben zu erstellen. Dennoch bildeten die genannten und eine Reihe weiterer Beispiele die Ausnahme als die Regel von digitalen geisteswissenschaftlichen Forschungsvorhaben im Zeitraum der 1970er bis 1990er Jahre. Sie sind eher als Avantgarde zu verstehen. Ein breiterer Trend hin zum Digitalen auch in den Geistes- und Kulturwissenschaften lässt sich erst seit ungefähr zehn Jahren erkennen (Neuroth 2012: 156). Einerseits lagen ab diesem Zeitraum immer mehr textbasierte Forschungsdaten, also Publikationen und Archivalien, elektronisch vor, andererseits trafen diese Entwicklungen auf eine immer größer werdende Zahl von digitalen Forschungsvorhaben und Initiativen, die sowohl von europäischer Ebene als auch durch das BMBF und die DFG gefördert wurden und dadurch eine allmähliche Öffnung und eine sich abzeichnende Erweiterung der geisteswissenschaftlichen Methodeninstrumentarien herbeiführten. Eine besondere Rolle nahmen hierbei die beiden durch das BMBF geförderten Forschungsinfrastrukturvorhaben CLARIN-D[16] und DARIAH-DE[17] ein, die aufgrund ihrer kollaborativen und auf Partizipation ausgerichteten konsortialen Strukturen digitale Werkzeuge entwickelten und Empfehlungen für die Nutzung von elektronischen Forschungsdaten konzipierten, die auf dezidierten fachwissenschaftlichen Anforderungen basierten.

Nichtsdestotrotz bestehen bis zum heutigen Tag drei Spannungsfelder, die bislang nicht aufgelöst wurden und insbesondere in den Geistes- und Kultur-

---

15 Siehe: http://www.tustep.uni-tuebingen.de (letzter Zugriff: 20. 10. 2017).
16 https://www.clarin-d.de/de/ (letzter Zugriff: 20. 10. 2017).
17 https://de.dariah.eu (letzter Zugriff: 20. 10. 2017).

wissenschaften, aber auch bei Kultureinrichtungen und Förderern in den kommenden Jahren thematisiert werden müssen:

1. Die an Geschwindigkeit permanent zunehmende Digitalisierung der Gesellschaft führt auch in den Geistes- und Kulturwissenschaften zu irreversiblen Veränderungen. Die Zunahme an elektronischen Publikationen und Archivalien, entweder digitalisiert oder sogar in maschinenlesbaren digitalen Formaten, führt zu veränderten und neuen Nutzungs- und Rezeptionsgewohnheiten. Dem werden sich Publikationsformen angleichen, da gedruckte Werke zukünftig nur noch in geringem Maße publiziert werden. Nimmt man die Veränderungen, die beispielsweise die Musikbranche in den vergangenen Jahren durchlebte – Phonograph (1877), Schellackplatte (1896), Vinyl-Schallplatte (1930), Kompaktkassette (1961), Compact Disc (1981), MP3 (1992), Filesharing (Mitte der 1990er), Musik-Streaming (seit 2005) – als Vorbild, so wird deutlich, dass einerseits die Abstände der innovativen Entwicklungen immer kürzer wurden, andererseits die Nutzungsmöglichkeiten sich allein in den vergangenen zehn Jahren grundsätzlich geändert haben. Musik ist seit diesem Zeitpunkt nicht nur zu jeder Zeit und an jedem Ort verfügbar, sondern das einzelne Individuum hat zugleich Zugriff auf den gesamten weltweiten Bestand. Die limitierenden Faktoren sind nicht mehr der Besitz eines einzelnen Stückes, sondern einzig allein der Netzzugang und ein entsprechendes Abo bei einem Streaming-Anbieter. Diese strukturellen Veränderungen betrafen nicht nur die individuellen Nutzungsmöglichkeiten, sondern hatten unmittelbare, wenn nicht sogar revolutionäre, Auswirkungen auf die Musikindustrie, die langjährig gewachsene Strukturen binnen kürzester Zeit anpassen mussten. Überträgt man diese Prozesse auf das Material, also beispielsweise Bücher, Quellen und Archivalien, das geisteswissenschaftlicher Forschung zugrunde liegt, so ist erkennbar, dass eine ‚undigitale' Forschung in Zukunft nicht mehr möglich sein wird – und zugleich, dass dieser revolutionäre strukturelle Wandel in ihrer Dimension noch aussteht. Gerade in diesen revolutionären medialen Veränderungen liegen aber auch zentrale Chancen. Einerseits können größere Buchbestände in ungeahnten Datenmengen digital erschlossen und analysiert werden, wie bereits Gregory Crane (2006) in seinem Aufsatz *What Do You Do with a Million Books?* postulierte; andererseits können neue Methoden für den Umgang und die Nutzung dieser digitalen Daten entwickelt werden. So unter anderem eine Erweiterung der Quellenkritik in der Geschichtswissenschaft für den Umgang mit digitalen Daten, um auch zukünftig grundsätzliche methodische Ansprüche, wie das Koselleck'sche Primat des „Vetorechts der Quellen" (Jordan 2010), an die neuen digitalen Anforderungen anpassen zu können. Dabei ist ganz zu schweigen von ge-

samtgesellschaftlichen Kontextualisierungen, die in einer digitalisierten und datengetriebenen Welt durch die Geisteswissenschaften erfolgen könnten, etwa der Verständlichmachung von Wissenschaft in der Öffentlichkeit oder den Risiken irregeleiteter Fakteninterpretationen (*Fake News*).

2. Die disziplinären und forschungspolitischen Diskurse um die Notwendigkeit der Etablierung von digitalen Forschungsinfrastrukturen sind in vielerlei Hinsicht immer noch von einem veralteten und zum Teil historisierenden Bild geprägt. Oftmals wird das klassische Bild einer Infrastruktur verwendet, wie beispielsweise das Autobahnnetz, die Versorgung mit Elektrizität, das Wassernetz oder das Bildungs- bzw. Gesundheitssystem. Hierbei handelt es sich um etablierte Infrastrukturen, die dafür Sorge tragen, dass spezifische Anforderungen einer gesamtgesellschaftlichen Versorgung erfüllt werden. Übertragen auf die geisteswissenschaftliche Forschung könnte man die klassische Versorgung mit gedruckter Literatur durch Bibliotheken, das Sammeln von Sammlungsgut durch Museen oder die Sicherung von Archivbeständen und deren Aufbereitung für die Forschung durch Archive synonym verwenden. Digitale Forschungsinfrastrukturen sind allerdings keine Bauten oder experimentelle Großgeräte, die einmal gekauft, funktionieren und ab und an repariert oder instandgesetzt werden müssen. Digitale Forschungsinfrastrukturen sind vielmals selbst Ausdruck der Innovation in der wissenschaftlichen Methodenentwicklung – und dies nicht nur in der Informatik oder den Informations-, Bibliotheks- und Archivwissenschaften – und befinden sich gerade deshalb in einem dauerhaften „Weiterentwicklungs- und Innovationszwang". Nur wenn ein kontinuierlicher Ausbau und Umbau für digitale Forschungsinfrastrukturen garantiert werden kann, können nachhaltige digitale Forschungsinfrastrukturen etabliert werden. Eine Aufgabe, die eben nicht nur für Geistes- und Kulturwissenschaften gilt, sondern auch in den Naturwissenschaften vorzufinden ist, wie das Beispiel CERN oder auch der Neubau des Forschungsschiffes *Polarstern*[18] zeigen. Insbesondere bei Softwarewerkzeugen, die Bestandteil digitaler Forschungsinfrastrukturen sind, besteht ein dauerhafter Entwicklungs- und Modernisierungsbedarf, hervorgerufen durch den technologischen Wandel und den immer kürzer werdenden Innovationszyklen bei den in Endgeräten eingesetzten Betriebssystemen und Schnittstellen. Dass selbst dies ein eigener Forschungsgegenstand ist und entsprechende Langzeitstudien fehlen, ist beispielsweise der seit 1994 jährlich veröffentlichten CHAOS-Studie der

18 https://www.awi.de/expedition/schiffe/polarstern.html (letzter Zugriff: 20. 10. 2017).

britischen Standish Group zu entnehmen, in der Erfolgs- und Misserfolgsfaktoren von IT-Projekten auf Basis von annähernd 40.000 Einzelprojekten analysiert werden (siehe z. B. The Standish Group 2014).

3. Zugleich besteht ein Spannungsfeld darin, dass zum jetzigen Zeitpunkt die Entwicklung und der Aufbau von digitalen Forschungsinfrastrukturen zwar durch die Forschungsförderer selbst als Forschungsvorhaben gehandhabt werden, aber zugleich für diese Forschungsprojekte bislang keine Möglichkeit bestand, Betriebsförderungen zu erhalten. Da der Betrieb von digitalen Forschungsinfrastrukturen, insbesondere in den Geistes- und Kulturwissenschaften, aufgrund der technologischen und inhaltlichen Komplexitäten nicht durch einzelne Einrichtungen getragen werden kann, müssen neben der Finanzierungsfrage zugleich neue Betriebsmodelle entwickelt werden. CLARIN-D und DARIAH-DE, die in diesen Fragen seit 2015 eng zusammenarbeiten, verfolgen beide einen kollaborativen und föderalen Ansatz, der eine Vielzahl unterschiedlicher Akteure mit unterschiedlichen Fähigkeiten und Kernkompetenzen in konsortialen Verbünden zusammenschließt. So werden insbesondere Schlüsselkompetenzen in daten-, forschungs-, tool- und lehrspezifischen Bereichen benötigt, die nur durch eine Kooperation von Universitäten, außeruniversitären Forschungseinrichtungen, Datenzentren, Bibliotheken, Archiven und auch klein- und mittelständischen Unternehmen aufgebaut und erfolgreich eingesetzt werden können.

Nach diesen Ausführungen zur speziellen Historizität digitaler Infrastrukturen im Allgemeinen kann nun DARIAH als eine der neuartigen Infrastrukturen für die Geisteswissenschaften vorgestellt werden, zunächst die Dachorganisation DARIAH-EU, die gemäß ihres Namens auf EU-Ebene, aber auch darüber hinaus agiert.

3 Eine kurze Geschichte von DARIAH-EU

DARIAH-EU wurde im Rahmen des ESFRI gegründet und tauchte erstmals 2006 in der ESFRI-Roadmap auf, als eines von sechs geistes- und sozialwissenschaftlichen Projekten (European Roadmap for Research Infrastructures 2006: 33). Innerhalb des ESFRI wurde eine Rechtsform entwickelt, die es den geförderten europäischen Forschungsverbünden ermöglicht, langfristig und finanziell stabil auf Basis eines europäischen Konsenses zu agieren. Diese Rechtsform trägt das Kürzel ERIC (*European Research Infrastructure Consortium*). Das DARIAH-

ERIC wurde nach langer Vorbereitung am 15. August 2014 von der Europäischen Kommission etabliert.

So wichtig dieser gesetzte politische Rahmen ist, näher an der Alltagspraxis ist man, wenn man DARIAH *bottom-up* erzählt. Direkt im DARIAH-Kontext forschen und arbeiten mehrere hundert Wissenschaftlerinnen und Wissenschaftler in ganz Europa, deren Institutionen (Universitäten, Bibliotheken, Forschungseinrichtungen, Zentren) wiederum direkte Kooperationspartner von DARIAH sein können. Die nächstgrößere Struktur sind die nationalen DARIAH-Verbünde, von denen DARIAH-DE, DARIAH-FR, DARIAH-IT und DARIAH-NL die größten sind.[19] Mittlerweile hat DARIAH 17 Mitgliedsländer, wobei diese nicht zwingend Mitglied der EU sein müssen, wie das Beispiel DARIAH-RS[20] (Serbien) zeigt. Hinzu kommen einzelne Institutionen, die Kooperationspartner sein können, ohne dass bereits deren Heimatländer volle DARIAH-Mitglieder sein müssen (momentan betrifft dies Institutionen in Finnland, Norwegen, Schweden, der Schweiz, Ungarn und in Großbritannien). Bevor wir einen Blick auf das Mitglied DARIAH-DE werfen, sei ein Blick auf die Organisationsstruktur von DARIAH-EU geworfen, um langsam von strukturellen zu inhaltlichen Gesichtspunkten überzuleiten.

Das Rückgrat von DARIAH bilden die vier Virtuellen Kompetenzzentren (VCCs), die sich um vier thematische Schwerpunkte gebildet haben, die für digital betriebene Geisteswissenschaften als entscheidend identifiziert wurden:

– VCC 1 (*e-Infrastructure*) beschäftigt sich mit dem technischen Fundament von DARIAH und sorgt für die Bereitstellung von Tools und Diensten für digital arbeitende Geisteswissenschaftlerinnen und -wissenschaftler.
– VCC 2 (*Research and Education Liaison*) dient als Schnittstelle zwischen den Infrastrukturangeboten und der Forschergemeinde und soll digitale Forschungspraktiken und -prozesse verstehen und fördern helfen. Ermöglicht wird dies etwa durch die Organisation von Workshops, bei denen junge Forscherinnen und Forscher mit den Methoden und Praktiken der digitalen Forschung bekannt gemacht werden, sowie durch die Abstimmung von Curricula in den *Digital Humanities* auf gesamteuropäischer Ebene.
– VCC 3 (*Scholarly Content Management*) beschäftigt sich mit den verschiedenen Stufen der Wissensgenerierung und soll sicherstellen, dass digitale Forschungsdaten zur Wiederverwendung bereitstehen, also deren Interoperabilität und Nachhaltigkeit gesichert sind, etwa durch die (Mit)Entwicklung von entsprechenden Standards.

19 https://www.dariah.eu/about/members-and-partners/; http://it.dariah.eu/sito/; https://www.clariah.nl/en/ (letzter Zugriff: 20. 10. 2017).
20 http://dariah.rs (letzter Zugriff: 20. 10. 2017).

– VCC 4 (*Advocacy, Impact and Outreach*) kümmert sich darum, dass DARIAH zum größtmöglichen Nutzen der Community arbeitet, und misst und bewertet dafür den Einfluss, die die Infrastruktur hat, sowie den Mehrwert, den sie bietet. VCC 4 versucht außerdem, das Einflussgebiet von DARIAH zu vergrößern, um gezielt Stakeholder zu erreichen.

Den direkten Draht zur Forschungscommunity stellen wiederum die DARIAH-Arbeitsgruppen her (*Working Groups*, WG), von denen es momentan 18 gibt. Sie sind einem oder mehreren VCCs zugeordnet, organisieren sich dabei aber selbst und reichen von Gruppen zu konkreten Forschungsfragen wie *Text and Data Analytics* über methodisch orientierte Gruppen wie *Digital Annotation* oder *Image Science and Media Art Research* bis hin zu eher integrativ ausgerichteten Gruppen wie *Training and Education* oder *Community Engagement*.

Schwenkt man den Blick von den WGs auf die *Governance*-Ebene, begegnet man verschiedenen Gremien, deren Funktionen in den Statuten des DARIAH-ERIC (2017) festgehalten sind:

– Die General Assembly (GA) versammelt VertreterInnen der Finanzmittelgeber, in den meisten Fällen RepräsentantInnen der nationalen Ministerien.
– Das Scientific Board (SB), das momentan aus zehn internationalen ExpertInnen aus dem Gebiet der Digital Humanities besteht, berät DARIAH in wissenschaftlichen und technischen Fragen.
– Das Board of Directors (BoD) ist das Exekutivorgan von DARIAH und besteht aus drei FachwissenschaftlerInnen, die für bis zu drei Jahre gewählt (und auch wiedergewählt) werden können.
– Das National Coordinator Committee (NCC) repräsentiert die KoordinatorInnen der nationalen DARIAHs auf gesamteuropäischer Ebene.
– Das Joint Research Committee (JRC) besteht aus den LeiterInnen der einzelnen VCCs und wird vom Chief Integration Officer (CIO) geführt.
– Das Senior Management Team (SMT) besteht aus der/dem Vorsitzenden plus StellvertreterIn des NCC und des Joint Research Committees (JRC).
– Das DARIAH Coordination Office (DCO) ist verantwortlich für Finanzen, Koordination und Kommunikation

Das Organisationsschema in Abbildung 3.3 kann nun zwar schematisch verdeutlichen, wie die einzelnen Gremien ineinandergreifen. Besser illustrieren lässt sich die Funktionsweise von DARIAH aber durch ein konkretes Beispiel, etwa am Entwurf und der (laufenden) Umsetzung des aktuellen DARIAH-Strategieplans. Schon in den Jahren 2011 und 2014 hatte es entsprechende Pläne gegeben, die zunächst darauf ausgerichtet waren, einen Überblick über die Wissenschaftslandschaft, in der sich DARIAH bewegt, zu gewinnen, ein *Mission Statement* zu entwerfen und *High-Level Principles* zu benennen. Basierend auf diesen Vorarbeiten zielt das aktuelle Papier, das Anfang 2017 als *living*

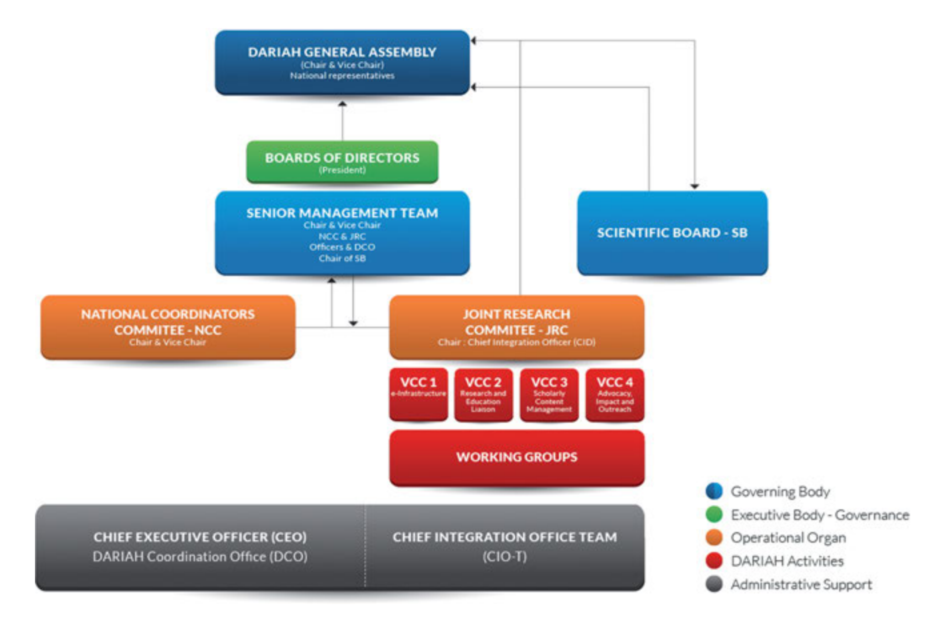


Abb. 3.3: Organigramm für DARIAH-EU (© DARIAH-EU 2017).

document ins Leben gerufen wurde, darauf ab, konkrete Ziele zu benennen und dafür Monitoring-Prozesse zu etablieren, die sich am Nutzen der zur Verfügung gestellten Dienste und Services für die Community ausrichten. Die 25 Aktionspunkte des Plans reichen zunächst bis ins Jahr 2019, sollen konkret aber den Betrieb von DARIAH über dieses Jahr hinaus gewährleisten.

Die Notwendigkeit eines aktuellen Strategieplans wurde, angestoßen von Laurent Romary (Präsident des Boards of Directors) zuerst im Januar 2017 bei einem Treffen des JRC und des SMT in Berlin diskutiert. Ein erster Entwurf, der von Jennifer Edmond (BoD), Sally Chambers (NCC) and Marianne Huang (JRC) verfasst wurde, zirkulierte ab Mitte Februar im SMT und im DCO, dann ab März im SB, JRC und NCC und wurde im April über die [dariah-all]-Mailingliste verteilt, zur Vorbereitung des Annual Meetings von DARIAH, das 2017 im Harnack-Haus in Berlin-Dahlem stattfand. Im Mai 2017 wurde das Papier in einer virtuellen Sitzung der General Assembly vorgestellt; eine finale Revision soll der General Assembly im November 2017 zur Bestätigung vorgelegt werden.

Die Beschreibung der Zirkulation dieses Papiers soll verdeutlichen, wie ausgreifend und aktiv Entscheidungsprozesse in einer Infrastruktur wie DARIAH organisiert werden, um alle Gremien, vor allem aber auch die (Repräsentanten der) Stakeholder einzubeziehen, gerade wenn es sich um ein paneuropäisches Infrastrukturprojekt wie DARIAH handelt.

4 Eine kurze Geschichte von DARIAH-DE

Seit das Bundesministerium für Bildung und Forschung (BMBF) 2011 das Pilotprojekt *Roadmap für Forschungsinfrastrukturen* initiiert hat, steht DARAH-DE als eine im Aufbau befindliche Forschungsinfrastruktur für die Geisteswissenschaften auf der nationalen Roadmap[21] – übrigens als eine von insgesamt zwei. Die andere Infrastruktur für die Geisteswissenschaften auf der Roadmap ist das bereits erwähnte CLARIN-D.

Die Förderung durch das BMBF hat DARIAH-DE bisher drei Projektlaufzeiten ermöglicht: die Aufbauphase 03/2011–02/2014, daran anschließend die Ausbau- und Konsolidierungsphase 03/2014–02/2016 und aktuell die Institutionalisierungsphase 03/2016–02/2019, der ab 03/2019 die Betriebsphase folgen soll. DARIAH-DE ist eine konsortial organisierte Initiative, deren 19 Partner deutschlandweit an den Standorten Bamberg, Berlin, Bonn, Darmstadt, Essen, Göttingen, Jülich, Karlsruhe, Mainz, München, Tübingen, Wolfenbüttel und Würzburg verteilt sind. Auch die Kompetenzen im Konsortium sind gut verteilt und ergänzen sich: Universitäten, Forschungseinrichtungen, Rechenzentren, Bibliotheken, Akademien der Wissenschaften, eine nichtstaatliche Organisation und ein kommerzieller Partner sorgen für die nötige fachwissenschaftliche, informationstechnische und informationswissenschaftliche Expertise, die Grundvoraussetzung für eine funktionierende Forschungsinfrastruktur ist.[22]

Strukturell war DARIAH-DE gerade in der Aufbauphase sehr eng an DARIAH-EU angelehnt, die vier Arbeitspakete *e-Infrastruktur* (AP 1), *Forschung und Lehre* (AP 2), *Forschungsdaten* (AP 3) und *DARIAH-DE Konsortium Management* (AP 4) waren analog zu den vier VCCs angelegt und sollten eine direkte Übertragbarkeit der erzielten Ergebnisse auf EU-Ebene (*in-kind contributions*) gewährleisten. Für die folgenden Projektphasen wurde eine flexiblere Cluster-Struktur gewählt, die es ermöglichte, neue Schwerpunkte als zusätzliche Cluster zu integrieren bzw. Cluster aufzulösen, wenn alle Inhalte erarbeitet worden sind. Damit vollzog DARIAH-DE organisatorisch einen wichtigen Schritt Richtung Institutionalisierung. Die Cluster sind thematisch ausgerichtet:

– Cluster 1 (*Begleitforschung*) achtet auf die Einhaltung von Nutzerfreundlichkeit und Erfolgskriterien bei der Anpassung bestehender sowie der

21 https://www.bmbf.de/pub/Roadmap_Forschungsinfrastrukturen.pdf (letzter Zugriff: 20. 10. 2017).

22 Einen Überblick über die Gesamtarchitektur, zur Governancestruktur, zu den einzelnen Forschungsthemen und dem Serviceangebot von DARIAH-DE findet sich in: Neuroth/Schmunk/Blümm/Rapp/Jannidis/Wintergrün/Schwardmann/Gietz (2016).

Integration und Entwicklung neuer Services und erarbeitet Publikations- sowie Disseminationsstrategien. Cluster 1 ist eng verzahnt mit VCC4.
- Cluster 2 (*eInfrastruktur*) schafft die technischen Voraussetzungen, um Dienste innerhalb von DARIAH-DE zuverlässig und dauerhaft zu betreiben. Es arbeitet direkt mit VCC1 zusammen.
- Cluster 3 (*Institutionalisierung von DARIAH-DE und Aufbau des DARIAH-DE Coordination Office*) kümmert sich um den organisatorischen und rechtlichen Aufbau des Coordination-Office und die Institutionalisierung für den dauerhaften Betrieb von DARIAH-DE. Es ist weiterhin verantwortlich für die Anbindung der fachwissenschaftlichen DH-Communitys sowie den Betrieb und weiteren Ausbau der DARIAH-DE Service Unit (DeISU).
- Cluster 4 (*Wissenschaftliche Sammlungen*) unterstützt Geisteswissenschaftlerinnen und -wissenschaftler bei der Erstellung und Nutzung von Wissenschaftlichen Sammlungen und Forschungsdaten und baut die DARIAH-DE Forschungsdaten-Föderationsarchitektur auf. Cluster 4 steht in Austausch mit VCC3.
- Cluster 5 (*Quantitative Datenanalyse*) erforscht die geisteswissenschaftliche Nutzung quantitativer Methoden und Verfahren wie Information Retrieval und Text Mining anhand von Use Cases zu biographischen Daten und zu Topic Modeling-Verfahren bei Textsammlungen.
- Cluster 6 (*Annotieren*, *analysieren*, *visualisieren*) entwickelt Strategien, um die Daten vorhandener Repositorien für Technologien des Semantic Web zu öffnen und somit nachhaltiger und effizienter nutzen zu können. Dies beinhaltet Use Cases zur Analyse und Visualisierung annotierter Forschungsdaten sowie den Austausch mit der Forschungscommunity. Cluster 5 und 6 sind an VCC2 angebunden.

Ein siebtes Cluster mit dem thematischen Schwerpunkt *Bilder und Objekte* ist in Planung. Diese Struktur ermöglicht es, dass DARIAH-DE als Forschungsinfrastruktur auch zukünftig neue Themenfelder aufnehmen und in Form von Clustern integrieren kann.

Genau wie DARIAH-EU weist auch DARIAH-DE einige zusätzliche unterstützende Strukturen und Einheiten auf (Abb. 3.4). Die Gesamtkoordination des Verbundprojekts erfolgt durch die Konsortialleitung. Daneben gibt es einige Arbeitsgruppen, die sich mit cluster- und institutionsübergreifenden Themen beschäftigen, z. B. der Aufnahme externer Dienste in die DARIAH-DE Forschungsinfrastruktur oder der Pflege des Projektportals.[23]

23 https://de.dariah.eu/ (letzter Zugriff: 20. 10. 2017).

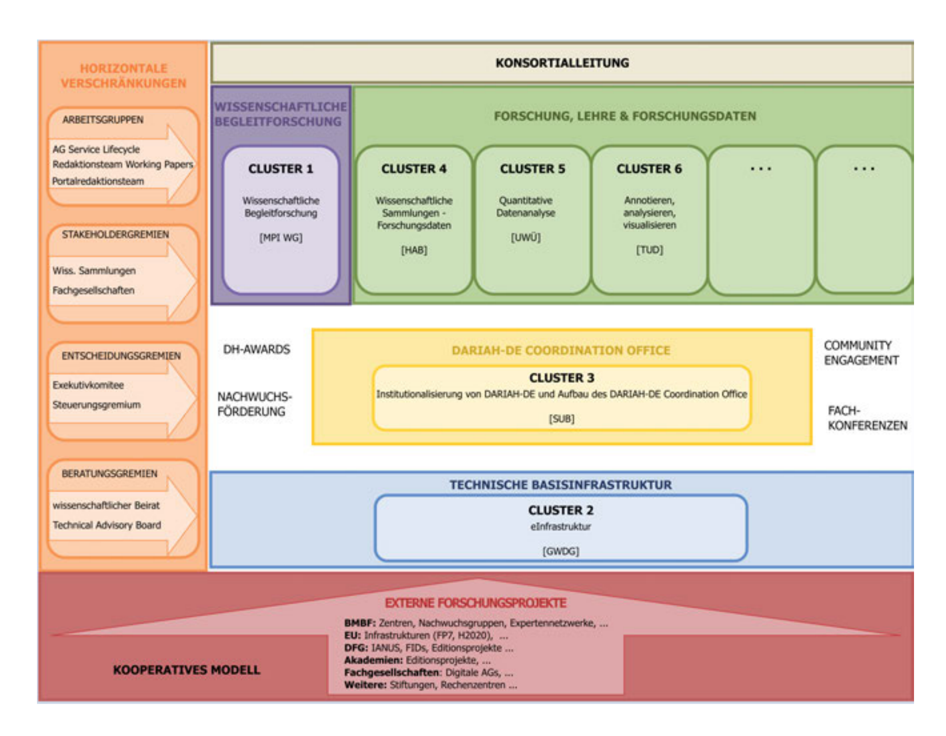


Abb. 3.4: Organigramm für DARIAH-DE (© DARIAH-DE 2016).

Fachgremien wie das Stakeholdergremium *Wissenschaftliche Sammlungen* versammeln externe Expertinnen und Experten zu einem speziellen Thema, die DARIAH-DE ihre Expertise zur Verfügung stellen und gemeinsam das Thema weiter entwickeln.

Zwei Entscheidungsgremien, das Steuerungsgremium als Organ der beteiligten Partnereinrichtungen und das Exekutiv-Komitee als Gremium der Clusterleitungen, monitoren den Stand der Entwicklungen und legen die Gesamtstrategie von DARIAH-DE fest. Der Wissenschaftliche Beirat[24] und der technische Beirat, das so genannte Technical Advisory Board[25], begleiten beratend die Entwicklung von DARIAH-DE. DARIAH-DE ist wie DARIAH-EU offen für Kooperationen und Austausch.

---

24 https://wiki.de.dariah.eu/display/publicde/Wissenschaftlicher+Beirat (letzter Zugriff: 20. 10. 2017).

25 https://wiki.de.dariah.eu/display/publicde/Technical+Advisory+Board (letzter Zugriff: 20. 10. 2017).

5 Die Rolle der Community

DARIAH-DE setzt das Konzept der „Architektur der Partizipation“ (*Architecture of participation*) (vgl. Blümm, Neuroth & Schmunk 2016), das im Kontext von DARIAH-EU entwickelt wurde, auf nationaler Ebene als kooperatives Modell um. Zentral dafür ist ein intensiver Austausch bzw. eine Verschränkung von Forschungsarbeiten und Entwicklungen von DARIAH-DE mit assoziierten Projekten bzw. Partnern im In- und Ausland. Durch das Zusammenwachsen mit TextGrid sind weitere Kooperationspartner hinzugekommen, womit das Zusammenwachsen von Communitys begünstigt wurde. So ist es gelungen, in den letzten Jahren eine breite Nutzerschaft aufzubauen und diese durch die verschiedenen Angebote der Forschungsinfrastruktur zu unterstützen. Gleichzeitig fließen Ideen, Anforderungen, aber auch Ergebnisse der an DARIAH-DE angedockten Forschungsprojekte, Initiativen und Verbünde in die weitere Entwicklung und Ausrichtung der Forschungsinfrastruktur maßgeblich ein.

Der Grad der Kooperation ist dabei variabel. Er reicht von einer Beteiligung im Konsortium durch Eigenmittel (z. B. Max Weber Stiftung) über eine Absichtserklärung (*Letter of Intent*) der Zusammenarbeit im Projektkontext bis hin zur reinen Nutzung von Diensten ohne näheren, direkten Austausch mit DARIAH-DE.

Durch diesen Ansatz wird maximale Flexibilität erreicht, die sicherstellen soll, dass Interessierte jeweils ihrer individuellen Bedarfe und Möglichkeiten gemäß die Angebote der Forschungsinfrastruktur nutzen können.

Abbildung 3.5 führt die Offenheit dieses Konzepts („Kooperatives Modell“) vor Augen. DARIAH-DE und TextGrid kooperieren mit rund 90 Projekten, Arbeitsgruppen, Initiativen und anderen Forschungsinfrastrukturen (Stand: Juli 2017), die an unterschiedlichen Institutionen – Universitäten, Bibliotheken, Archiven, Akademien und außeruniversitären Forschungseinrichtungen – verortet sind, sowie teilweise mit den Institutionen selbst. Vorhaben und Projekte, die sich unmittelbar an der Weiterentwicklung von Komponenten von DARIAH-DE und TextGrid beteiligen, sind in den beiden Kreisen hinterlegt. Mit diesen werden in enger Abstimmung Entwicklungen in bestimmten Bereichen gemeinsam vorangetrieben. Alle anderen in der Grafik angegebenen Forschungsvorhaben, Drittmittelprojekte und sonstige Initiativen sind lockerer mit den Infrastrukturen verbunden. Sie nutzen beispielsweise einzelne Dienste, entwickeln Tools weiter oder erzeugen Forschungsdaten, die an DARIAH-DE zurückgespielt und so der Community zugänglich gemacht werden. Eine große Zahl an Projekten/Initiativen greift zudem auf Angebote sowohl von DARIAH-DE als auch TextGrid zurück.

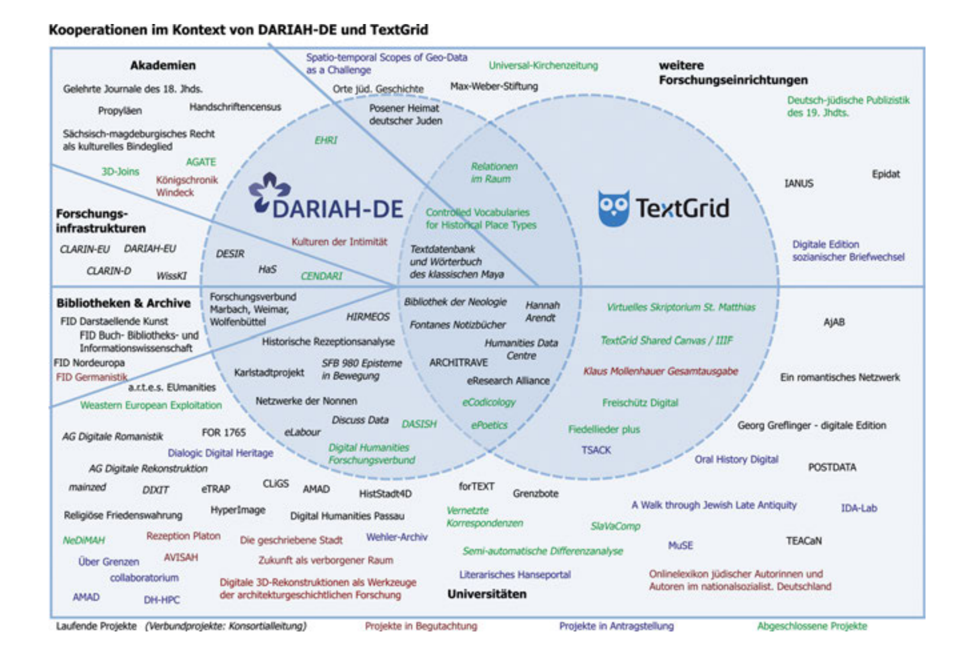


Abb. 3.5: Kooperatives Modell von DARIAH-DE (© DARIAH-DE 2017).

Rund 59 Prozent der Kooperationen sind laufende Vorhaben (schwarz eingefärbt), rund 10 Prozent der Vorhaben befinden sich derzeit in Begutachtungsprozessen (rot eingefärbt) und zirka 9 Prozent der Vorhaben stehen gerade im Antragsstadium bzw. kurz vor der Einreichung. Eine wachsende Zahl von Kooperationsprojekten (rund 22 Prozent) ist abgeschlossen und sichert ihre Projektergebnisse über DARIAH-DE und TextGrid. Es ist davon auszugehen, dass die Zahl in den nächsten Jahren weiter ansteigen wird, so dass DARIAH-DE dies in seinen Geschäftsmodellen berücksichtigen muss.[26]

6 Designprinzipien für digitale Forschungsinfrastrukturen

DARIAH-DE und DARIAH-EU zeigen Modelle dafür auf, wie in den Geistes- und Kulturwissenschaften eine digitale Forschungsinfrastruktur gebaut, betrieben

26 Einen Überblick zu den derzeit geführten Diskursen zu Nachhaltigkeit, siehe in: Blümm/Schmunk/Gietz/Horstmann/Hütter (2016).

und organisiert werden kann. Die besondere Bedeutung dieser Modelle liegt darin, dass die drei oben angesprochenen Herausforderungen adressiert und praktisch aufgezeigt werden: 1. die zunehmende Digitalisierung und das Datenwachstum, 2. die dynamische Gestalt von digitalen Forschungsinfrastrukturen (im Unterschied zu Bauten und Großgeräten) und 3. die (bisher schwach ausgebildeten) Betriebsförderungsmodelle. Hinzu kommt, dass im Unterschied zu digitalen Forschungsinfrastrukturen in den Natur- und Technikwissenschaften mit ihren Verbünden für Großgeräte und Experimentalumgebungen die Forschungsbasis von DARIAH häufig in der Einzelforschung liegt. Besonders deutlich wird die Ausbildung von geeigneten Designprinzipien in der Betrachtung der Situation in Deutschland (siehe Abb. 3.5).

Zum einen besteht das *unabdingbare Primat einer interoperablen Infrastruktur*, durch das Infrastrukturkomponenten wie Speicher, Authentifizierung, Software-Werkzeuge und Datenstrukturen projektweise angepasst werden, um im Ergebnis Daten und Quellen für die Forschung von hoher Diversität zu erzeugen, die jedoch der Forschung disziplinenübergreifend zur Verfügung gestellt werden können. Zum anderen wurde eine *Architektur der Partizipation* etabliert, in deren Rahmen universitäre und außeruniversitäre Forschung, unabhängig vom Organisationstyp und Forschungsfeld, stattfinden kann. Dies demonstriert, dass die Forschungsfrage entscheidend ist und nicht die Organisationsform oder das Förderformat, was besonders an der Durchdringung verschiedener Forschungsförderer abgelesen werden kann. Diese freie und eigenbestimmte „Bestenauslese“ von Projektpartnern für die Bearbeitung einer Forschungsfrage repräsentiert nicht nur gelebte Wissenschaftsfreiheit, sondern zeigt eben auch, dass die Selbstorganisationsprinzipien der Wissenschaft und die Formierung von gemeinsamen, funktionierenden Informationsinfrastrukturen nicht im Widerspruch stehen.

Die Designprinzipien des *unabdingbaren Primats einer interoperablen Infrastruktur* und einer *Architektur der Partizipation* können sogar als Blaupause für die Natur- und Technikwissenschaften wertvoll sein, in denen es zwar teils gelungen ist – analog zu Bibliotheken und Archiven für das Schriftgut – Informationsinfrastrukturen um experimentelle Großgeräte zu formieren (etwa im CERN), in denen jedoch kaum Informationsinfrastrukturen für die Einzelforschung existieren. Eingedenk des Umstandes, dass wissenschaftliche Innovationen dort entstehen, wo über Einzelforschung in bisher kaum betrachteten Feldern agile, interdisziplinäre, länder- und institutionenübergreifende Verbünde entstehen, sind Modelle für Informationsinfrastrukturen für diese Rahmenbedingungen eine entscheidende Zukunftsaufgabe für die Wettbewerbsfähigkeit von Wissenschaftsstandorten.

Literatur

Blanke, Tobias, Michael Bryant , Marc Hedges, Andreas Aschenbrenner, Mike Priddy (2011): Preparing DARIAH. *IEEE 7th International Conference on E-Science (e-Science)*, 158–165.

Blümm, Mirjam, Stefan E. Funk & Sibylle Söring (2015): Die Infrastruktur-Angebote von DARIAH-DE und TextGrid. *Information. Wissenschaft & Praxis* 66(5–6), 304–312.

Blümm, Mirjam, Heike Neuroth & Stefan Schmunk (2016): DARIAH-DE – Architecture of participation. *Bibliothek – Forschung und Praxis* 40, Heft 2, 165–171, doi:10.1515/bfp-2016-0026.

Blümm, Mirjam, Stefan Schmunk, Peter Gietz, Wolfram Horstmann & Heiko Hütter (2016): Vom Projekt zum Betrieb: Die Organisation einer nachhaltigen Infrastruktur für die Geisteswissenschaften DARIAH-DE. *ABI Technik* 36 (1), 10–23. doi:10.1515/abitech-2016-0011.

Crane, Gregory (2006): What do you do with a million books? *D-Lib Magazine* 12 (3). doi:10.1045/march2006-crane (letzter Zugriff: 20. 10. 2017).

DARIAH-ERIC (2017): *Statutes* (April 2017). http://www.dariah.eu/fileadmin/Documents/statutes/170405_DARIAH_ERIC_Statutes.pdf (letzter Zugriff: 20. 10. 2017).

Europäisches Parlament (Hg.) (2016): *Der Europäische Forschungsraum. Ein Konzept in der Entwicklung, eine Herausforderung bei der Umsetzung*. Straßburg: Wissenschaftlicher Dienst des Europäischen Parlaments. doi:10.2861/48523.

European Roadmap for Research Infrastructures (2006): *Report 2006*. Luxembourg: Office for Official Publications of the European Communities. https://ec.europa.eu/research/infrastructures/pdf/esfri/esfri_roadmap/roadmap_2006/esfri_roadmap_2006_en.pdf (letzter Zugriff: 20. 10. 2017).

Grotum, Thomas (2004): *Das Digitale Archiv. Aufbau und Auswertung am Beispiel der Geschichte des Konzentrationslagers Auschwitz*. Frankfurt a. M.: Campus.

Jordan, Stefan (2010): Vetorecht der Quellen. Version 1.0. In *Docupedia-Zeitgeschichte* 11. Februar 2010. http://dx.doi.org/10.14765/zzf.dok.2.570.v1 (letzter Zugriff: 20. 10. 2017).

Mak, Geert (1999): *Wie Gott verschwand aus Jorwerd. Der Untergang des Dorfes in Europa*. Berlin: btb Verlag.

Neuroth, Heike, Stefan Schmunk, Mirjam Blümm, Andrea Rapp, Fotis Jannidis, Dirk Wintergrün, Ulrich Schwardmann & Peter Gietz (Hrsg.) (2016): *Bibliothek Forschung und Praxis* 40 (2), Sonderheft: Digitalität in den Geistes- und Kulturwissenschaften am Beispiel der digitalen Forschungsinfrastruktur DARIAH-DE. doi:10.1515/bfp-2016-0019.

Neuroth, Heike (2012): DARIAH-DE. Forschungsinfrastrukturen für die eHumanities. *BIS. Das Magazin der Bibliotheken in Sachsen*, 156–158.

The Standish Group (Hrsg.) (2014): Report 2014. London: The Standish Group. https://www.projectsmart.co.uk/white-papers/chaos-report.pdf (letzter Zugriff: 20. 10. 2017).

Wikipedia (2017): Artikel „Infrastruktur“. *Die freie Enzyklopädie*. Bearbeitungsstand: 15. Juni 2017. https://de.wikipedia.org/w/index.php?title=Infrastruktur&oldid=166415027 (letzter Zugriff: 20. 10. 2017).

Wissenschaftsrat (Hrsg.) (2011): *Empfehlungen zu Forschungsinfrastrukturen in den Geistes- und Sozialwissenschaften*. Berlin. https://www.wissenschaftsrat.de/download/archiv/10465-11.pdf (letzter Zugriff: 20. 10. 2017).

Karlheinz Mörth und Tanja Wissik

4 Digitale Sprachressourcen in Österreich

Abstract: Der Beitrag skizziert rezente Entwicklungen im Zusammenhang mit dem Aufbau digitaler Infrastrukturen für text- und sprachbasierte Forschung in Österreich und demonstriert anhand ausgewählter Beispiele das weite Spektrum unterschiedlicher digitaler Sprachressourcen, die in den letzten Jahren entstanden sind.

Keywords: Digitalisierung, Forschungswerkzeuge, Sprachkorpora, Vernetzung, Wörterbücher

1 Vorbemerkungen

Die pervasive Natur der „digitalen Revolution" lässt digitale Sprachressourcen in vielerlei Gestalt in unseren Alltag vordringen. Dies trifft natürlich auch auf die Forschung zu, insbesondere auf alle text- und sprachorientierten Disziplinen. Was unter digitalen Sprachressourcen genau zu verstehen ist, wurde in der Literatur bislang auf unterschiedliche Weise abgehandelt. Es lassen sich zwei Ansätze unterscheiden: ein eher enger definierter Begriff versteht unter Sprachressourcen vor allem Sprachdaten, digitale Texte, Lexika und Metainformationen über diese.

Trippel, Declerck & Heid (2005) definieren Sprachressourcen als

> eine Klasse heterogener Informationen, die Gegenstand von Linguistik und Sprachtechnologie sind, aber auch in Anwendungskontexten wie Übersetzung und Lexikonentwicklung gefragt sind. Dazu gehören Textkorpora, Lexika, Daten gesprochener Sprache, aber auch Annotationsrichtlinien und -verfahren. (Trippel, Declerck & Heid 2005: 17)

Manche Definitionen gehen von einem weiteren Verständnis aus, welches das gesamte Repertoire an digitalen Werkzeugen, die zur Manipulation von Sprachdaten dienen, einschließt:

Karlheinz Mörth, Österreichische Akademie der Wissenschaften, Sonnenfelsgasse 19, A-1010 Wien, E-Mail: karlheinz.mörth@oeaw.ac.at
Tanja Wissik, Österreichische Akademie der Wissenschaften, Sonnenfelsgasse 19, A-1010 Wien, E-Mail: tanja.wissik@oeaw.ac.at

https://doi.org/10.1515/9783110538663-005

> A Language Resource is any physical or digital item that is a product of language documentation, description, or development, or is a tool that specifically supports the creation and use of such products. (Simons & Bird 2008: 88)

Bereits Trippel, Declerck & Heid (2005) sprachen unterschiedliche Anwendungskontexte an. Darüber hinaus nutzen unterschiedliche kultur- und sozialwissenschaftliche Disziplinen ebenso Sprachressourcen, auch wenn sie nicht immer dieselbe Terminologie verwenden oder oft nicht explizit von Sprachressourcen sprechen. Eine Umfrage, die im Rahmen des Forschungsinfrastrukturkonsortiums CLARIN (*Common Language Resources and Technology Infrastructure*) durchgeführt wurde, hat gezeigt, dass die Nutzergruppe von Sprachressourcen aus den unterschiedlichsten Bereichen wie zum Beispiel der Linguistik, der Computerlinguistik, den Literaturwissenschaften, der Altphilologie, der Geschichte, den Medienwissenschaften, der Ethnologie, der Lexikographie und den Politikwissenschaften, um nur einige zu nennen, kommt (vgl. Wynne 2015). Daher werden wir in unserem Beitrag versuchen, Sprachressourcen aus teilweise sehr unterschiedlichen Disziplinen zu präsentieren. Das diesem Papier zugrunde liegende Verständnis von Sprachressourcen folgt dem zweiten, weiteren Ansatz und umfasst sowohl Daten als auch relevante digitale Werkzeuge.

An dieser Stelle sei darauf hingewiesen, dass der Umfang dieses Beitrags nur die ausgewählte Nennung von Projekten und Ressourcen zulässt. Wir konzentrieren uns dabei auf digitale Sprachressourcen, die rezent im akademischen Raum entstanden sind oder dort Verwendung finden.

2 Infrastrukturen

2.1 Von FLaReNet, META-NET und META-SHARE zu CLARIN-ERIC

Zeitlich begrenzte Projekte stellen bekanntlich ein probates Mittel dar, die Forschung voranzutreiben; im Hinblick auf Infrastrukturen sind sie allerdings weniger geeignet, Nachhaltigkeit sicherzustellen. In den Geisteswissenschaften kamen und gingen in den letzten Jahren zahllose Projekte, die an digitalen Infrastrukturen arbeiteten. ARIADNE, CENDARI, CHARISMA, DASISH, EHRI und NEDIMAH sind nur einige Beispiele dafür. Für digitale Sprachressourcen waren in diesem Zusammenhang insbesondere FLaReNet, META-NET und META-SHARE von Relevanz, in denen in Österreich vor allem die Universität Wien engagiert war. Hier können diese Aktivitäten als einer der Ausgangspunkte für Entwicklungen gesehen werden, an deren Endpunkt etliche der noch zu besprechenden Projekte, die Gründung des *Austrian Centre for Digital*

Humanities an der Österreichischen Akademie der Wissenschaften (ACDH-OeAW) und die Konstituierung des österreichischen Zweiges von CLARIN und DARIAH stehen. Auf europäischer Ebene stellte die Einrichtung des neuen Rechtskonstrukts der europäischen Forschungsinfrastrukturkonsortien, der ERICs (*European Research Infrastructure Consortia*), einen Versuch dar, mehr Nachhaltigkeit und längerfristige Operabilität in die Entwicklung von Infrastrukturen zu bringen. Seit 2012 ist CLARIN offiziell ein ERIC und darum bemüht, in den Mitgliedstaaten ein breites Spektrum an Technologien und Services für digitale Sprachdaten und Tools verfügbar zu machen. Eine der ersten österreichischen CLARIN-Aktivitäten war eine Umfrage zu digitalen Sprachressourcen, die im Jahr 2009 durchgeführt wurde. Diese Erhebung sollte den Status Quo in Österreich feststellen und konkrete Vorschläge für die Vorbereitungs- und Aufbauphase des Netzwerkes in Österreich erarbeiten (Wissik & Budin 2010).

2.2 CLARIAH-AT und dha

In Österreich waren die aus der ESFRI-Roadmap hervorgegangen Aktivitäten von CLARIN und DARIAH (*Digital Research Infrastructure for the Arts and Humanities*) von allem Anfang an personell und institutionell stark verschränkt und wurden als Einheit konzipiert, was ein Höchstmaß an Synergien und Effizienz sicherstellen sollte. Während die Universität Wien bis 2013 die koordinierende Institution war, übernahm ab 2014 die Österreichische Akademie der Wissenschaften (ÖAW) im Auftrag des Bundesministeriums für Wissenschaft, Forschung und Wirtschaft (BMWFW) diese Aufgabe. Seit damals agiert die Kerngruppe unter dem Namen CLARIAH-AT. Nach außen tritt sie als virtuelles Netzwerk unter dem Namen *Digital Humanities Austria* (dha[1]) auf, das auch eine jährliche Konferenz veranstaltet.

Eine der Maßnahmen, die auch im Rahmen der österreichischen CLARIN- und DARIAH-Aktivitäten angesiedelt waren, ist die von der ÖAW organisierte und aus Mitteln der Akademie, des BMWFW und der österreichischen Nationalstiftung für Forschung, Technologie und Entwicklung finanzierte *go!digital*-Förderschiene für innovative digitale Projekte. Bisher konnten in zwei Durchgängen insgesamt 17 Projekte gefördert werden, die sich durch einen besonders hohen Grad an methodologischer Innovation auszeichneten und dazu angetan waren, nachhaltig in die österreichische Forschungslandschaft hinein zu wirken.

1 www.digital-humanities.at (letzter Zugriff: 24. 10. 2017).

Auf der institutionellen Ebene ist auch die Gründung des *Austrian Centre for Digital Humanities* der ÖAW (ACDH-OeAW) im Jahre 2015 zu erwähnen, an dem digitale Sprach- und Literaturwissenschaften eine wichtige Rolle spielen. Text-technologische Verfahren und digitale Sprachressourcen nehmen im Portfolio des Instituts einen zentralen Platz ein. Dies und die Rolle des Instituts in CLARIAH-AT sind die Hauptgründe dafür, dass es in den folgenden Ausführungen eine prominente Stellung einnimmt.

2.3 Werkzeuge und Dienste

Wie eingangs bereits ausgeführt, gehen wir von einem weiten Begriff der Sprachressourcen aus, der auch digitale Werkzeuge zur Erzeugung, Modifikation, Präservation und Analyse digitaler Sprachdaten einschließt.

2.3.1 GAMS, PHAIDRA, CCV: Langzeitarchivierung für Forschungsdaten

Wichtige Eckpfeiler einer modernen digitalen Forschungsinfrastruktur sind Repositorien zur Langzeitspeicherung digitaler Forschungsdaten. Hier befinden wir uns vielerorts erst in einer Aufbauphase, bestehende Strukturen sind kaum in der Lage, den in den letzten Jahren gewaltig gestiegenen Bedarf zu decken. Diskussionen drehen sich um den Gegensatz zwischen distribuierten, kleinen, institutionellen Lösungen und großen, zentral angelegten Datenzentren.

Unter den institutionellen Repositorien in Österreich seien nur drei hervorgehoben, die im geisteswissenschaftlichen Umfeld eine besondere Rolle spielen. Das erste ist das *Geisteswissenschaftliche Asset Management System* (GAMS), welches am Zentrum für Informationsmodellierung – *Austrian Centre for Digital Humanities* an der Karl-Franzens-Universität Graz seit mehreren Jahren entwickelt und betrieben wird. Es ist dies ein OAIS-konformes FEDORA-basiertes System, das auf eine weitgehend XML-basierte Content-Strategie und zahlreiche systeminhärente Funktionalitäten zur Verwaltung und Publikation der digitalen Daten baut (GAMS 2017). Die an GAMS arbeitende Gruppe ist auch stark in CLARIAH-AT und DARIAH.EU involviert.

Das PHAIDRA-Repositorium (*Permanent Hosting, Archiving and Indexing of Digital Resources and Assets*) wird seit einigen Jahren an der Universität Wien entwickelt und mittlerweile an über einem Dutzend weiterer Institutionen im In- und Ausland verwendet (Blumesberger 2010). Wie GAMS baut es auf FEDORA auf.

Zuletzt sei noch das erste offizielle CLARIN-Zentrum in Österreich, das *CLARIN Centre Vienna* (CCV), erwähnt, welches als Repositorium für digitale Sprachressourcen seit 2014 vom ACDH-OeAW betrieben wird.

2.3.2 ACDH-Tools: VLE, *tokenEditor*

Am ACDH-OeAW wurde in den letzten Jahren im Rahmen zahlreicher Projekte auch an digitalen Tools und Services gearbeitet. Beispielhaft seien der *Viennese Lexicographic Editor* (VLE) und der *tokenEditor* erwähnt.

VLE ist ein XML-Editor, der zum kollaborativen Arbeiten an lexikographischen Daten dient. Er wurde im Rahmen mehrerer lexikographischer Projekte entwickelt und verfügt über eine Reihe spezialisierter Funktionen, die typischerweise bei der Bearbeitung lexikographischer Daten Verwendung finden. Grundsätzlich für jede Art von XML-Daten verwendbar sind viele Funktionen besonders auf das Datenmodell der TEI hin optimiert. Aus den unterschiedlichen Projekten heraus entstanden mehrere spezielle Module: so verfügt VLE über einen integrierten Book Reader und einen Internetbrowser, die den direkten Zugriff auf externe Quellen (Corpora, Wörterbücher etc.) ermöglichen. Rezente Versionen haben ein Modul, mit dessen Hilfe man in komfortabler Art und Weise einen Webauftritt für sein Wörterbuch erstellen kann, und ein Modul namens *tokenEditor*, welches es ermöglicht, Textcorpora komfortabel mit lexikalischen Informationen anzureichern.

Der *tokenEditor* des *VLE* war eine Vorform des *webbasierten tokenEditors*, der über die ACDH-OeAW Website[2] verfügbar ist und in zahlreichen Projekten zur manuellen Nachbearbeitung automatisch erzeugter Annotationen wie POS-Tags oder Lemma eingesetzt wird.

2.3.3 *Transkribus*: Texterkennung der nächsten Generation

Transkribus ist eine institutionell in Innsbruck verortete Initiative, die eine umfassende Plattform zur Erkennung und Transkription historischer Dokumente entwickelt hat. Die Plattform besteht aus drei wesentlichen Modulen: einem Expertenprogramm, einem Webinterface[3] und verschiedenen Diensten, die auf den Servern der Universität Innsbruck eingerichtet wurden. *Transkribus* stellt eine Reihe von Werkzeugen und Dienste für die automatisierte Erfassung von

2 https://www.oeaw.ac.at/acdh/tools/tokeneditor/ (letzter Zugriff: 24. 10. 2017).
3 http://transkribus.eu/ (letzter Zugriff: 24. 10. 2017).

Dokumenten zur Verfügung, darunter eine computergestützte Handschriftenerkennung, Bilderkennung und Strukturerkennung. Die Plattform ist für alle Benutzer frei zugänglich (Transkribus 2016). Die Entwickler der Plattform sind auch intensiv an CLARIAH-AT beteiligt.

3 Texte und Corpora

Die folgende Zusammenstellung unterschiedlicher Einzelprojekte und Ressourcen soll die weite Palette und große disziplinäre Vielfalt, in der moderne Corpustechnologie entsteht und zur Anwendung kommt, sichtbar machen.

3.1 *Vienna-Oxford International Corpus of English* (VOICE)

Eine besonders bemerkenswerte strukturierte Sammlung digitaler Sprachdaten stellt das VOICE-Corpus dar, das erste maschinenlesbare Corpus, das der Erforschung von Englisch als Lingua franca gewidmet ist. Es wurde am Institut für Anglistik der Universität Wien erstellt, umfasst 151 Sprachereignisse von 1.300 Sprecherinnen und Sprechern mit 50 unterschiedlichen Erstsprachen. Mit einem Umfang von einer Million Token ist das seit 2009 online verfügbare, in TEI-P5 kodierte Corpus eine Ressource, die in zahlreichen einschlägigen Untersuchungen Verwendung gefunden hat (Breiteneder et al. 2009; Osimk-Teasdale 2013).

3.2 *BAnca Dati dell'Italiano Parlato* (BADIP)

Das romanistische Institut der Karl-Franzens-Universität Graz bietet mit annähernd einer halben Million Wörtern seit vielen Jahren Zugang zu einem der größten Corpora des gesprochenen Italienisch. Das LIP (*Lessico di frequenza dell'italiano parlato*) Corpus enthält ca. 57 Stunden Gesprächsaufnahmen von bi-direktionaler Face-To-Face-Kommunikation mit freiem Sprecherwechsel (Gespräche zu Hause, in der Arbeit, in der Schule), bi-direktionaler computervermittelter Kommunikation mit freiem Sprecherwechsel (z. B. Telefongespräche), bi-direktionaler Face-To-Face-Kommunikation mit reglementiertem Sprecherwechsel (z. B. Parlamentssitzungen, Prüfungen, Interviews), monodirektionale Kommunikation (z. B. wissenschaftliche Vorträge, politische Reden) sowie Radio- und Fernseh-Kommunikation (z. B. Radio- und Fernseh-

sendungen). Das transkribierte Corpus (ca. 490.000 Token) ist reich annotiert (Bellini & Schneider 2006), das Interface[4] erlaubt auch, nach Textsorten und Orten zu suchen.

3.3 Speech-Corpora des Instituts für Signalverarbeitung und Sprachkommunikation

Umfangreiche Speech-Corpora wurden in den letzten Jahren auch an der Technischen Universität Graz am Institut für Signalverarbeitung und Sprachkommunikation[5] erstellt, wie z. B. das *Austrian German Multi-Sensor Corpus* (AMISCO) oder das *Graz Corpus of Read And Spontaneous Speech* (GRASS).

Das AMISCO Corpus[6] beinhaltet die Audio- und Video-Aufnahmen (ca. 8,2 Stunden) und orthographische Transkriptionen (ca. 53.000 Token) von 24 sich im Raum bewegenden und nicht bewegenden Sprechern und Sprecherinnen (Pessentheiner, Pichler & Hagmüller 2016). Das GRASS Corpus beinhaltet Aufzeichnungen und orthographische Transkriptionen von spontaner Kommunikation sowie Leseaufgaben von 38 Sprecherinnen und Sprechern, die in Österreich geboren und aufgewachsen sind (Schuppler et al. 2014).

3.4 Digitale Sammlung der Grazer Linguistischen Slawistik (*Gralis*-Korpus)

Das *Gralis*-Korpus beinhaltet eine mehrsprachige und multifunktionale Sammlung von digitalen Texten, Audio-, Video-, TV- und anderen Daten, die für linguistische Untersuchungen zu slawischen Sprachen gesammelt und aufbereitet wurden. Es ist ein Projekt des slawistischen Instituts der Karl-Franzens-Universität Graz und wie auch das BADIP-Corpus durch eine enge Kooperation mit dem Zentrum für Informationsmodellierung in den Geisteswissenschaften möglich geworden. Das *Gralis*-Korpus[7] ist so angelegt, dass es als ein-, aber auch als mehrsprachiges Corpus (Parallelisierung von mindestens zwei slawischen Sprachen oder Parallelisierung einer slawischen Sprache und Deutsch)

---

4 http://badip.uni-graz.at/en/ (letzter Zugriff: 23. 10. 2017).

5 Signal Processing and Speech Communication Laboratory (SPSC Lab) https://www.spsc.tugraz.at// (letzter Zugriff:20. 11. 2017).

6 https://www.spsc.tugraz.at/tools/amisco (letzter Zugriff: 23. 10. 2017).

7 http://www-gewi.uni-graz.at/gralis/korpusarium/gralis_korpus.html (letzter Zugriff: 24. 10. 2017).

eingesetzt werden kann. Es besteht aus zwei großen Teilen, dem G*ralis Speech-Korpus* und dem *Gralis Text-Korpus*. Das *Gralis Speech-Korpus* besteht aus transkribierten Audioaufnahmen während das *Gralis Text-Korpus* neben parallelen Textcorpora auch Corpora der Werke von Literaten wie etwa Ivo Andrić, Zoran Živković u. a. enthält (Tošović 2017).

3.5 *Austrian Media Corpus* (AMC)

Das *Austrian Media Corpus* (AMC) ist Österreichs umfangreichstes linguistisch aufbereitetes digitales Corpus, das aus einer Kooperation zwischen der ÖAW und der Austria Press Agency (APA) hervorgegangen ist. Das AMC umfasst unterschiedliche journalistische Textsorten von 1986 bis heute (regionale und überregionale Tageszeitungen, Magazine, Pressemitteilungen, transkribierte Fernsehinterviews u. ä. m.) (vgl. Ransmayer, Mörth & Ďurčo 2013). Beim AMC handelt es sich um ein Monitorcorpus, welches laufend erweitert wird und aktuell 40 Millionen Artikel aus 53 Medien (Zeitungen, Magazine, TV-Transkripte), in Summe an die 10 Milliarden Token, beinhaltet. Die Rohdaten wurden mit linguistischen Basisannotationen versehen und in eine Infrastruktur überführt, die diese auch im internationalen Vergleich sehr umfängliche Sprachressource für linguistische Auswertungen zugänglich macht. Die Analysewerkzeuge ermöglichen neben der Erzeugung klassischer Konkordanzlisten auch Abfragen mittels morphosyntaktischer Muster, die Ermittlung von Wortähnlichkeiten und Erstellung von Konkordanzprofilen. Durch die verfügbaren Metadaten ist es möglich, die Abfrage auf spezifische Medien oder Ressorts sowie bestimmte Zeiträume oder Regionen zu beschränken und somit adhoc Subcorpora für die jeweiligen Analysen zu erstellen. Anwendungsbeispiele reichen von klassischen lexikographischen und linguistischen Fragestellungen (Ransmayr et al. 2016; Werner 2017) bis hin zu Forschungsfragen, die an der Schnittstelle zwischen Wirtschaftswissenschaften und computerbasierter Sprachanalyse liegen (Atz & Gerstbauer 2017).

3.6 *Transbank*

Das gerade in Angriff genommene Projekt *Transbank* ist am Zentrum für Translationswissenschaften der Universität Innsbruck angesiedelt und zielt darauf ab, eine großes, offenes und erweiterbares Corpus von übersetzten Texten und ihren Originalen aufzubauen. Die Daten werden auf Satzebene aligniert und mit Metadaten versehen, die auch translationsspezifische Informationen enthalten (Lusicky et al. 2017). Das Projekt wird eine Sprachressource schaffen,

die Daten, die fragmentiert an verschiedenen Orten bereits vorhanden sind, zusammenführt und für unterschiedliche Nutzergruppen und Forschungsfragen frei zugänglich macht (Transbank 2017).

3.7 INPUT und ADS: Kindersprache und Spontansprachcorpora

In die Reihe an Spezialsammlungen gehören auch die Kindersprachcorpora, die in den letzten zwei Jahrzehnten aus den Arbeiten der Arbeitsgruppe für Komparative Psycholinguistik am Institut für Sprachwissenschaft der Universität Wien hervorgegangen sind. Neben longitudinalen Spontansprachdaten von Wiener Kindern ist besonders das INPUT-Corpus (*Investigating Parental and Other Caretakers' Utterances to Kindergarten Children*) zu erwähnen, welches Aufnahmen von 61 Wiener Kindern im Alter von 3 bis 4,5 Jahren und ihrer Bezugspersonen enthält und in der Zeit von 2012 bis 2016 erstellt wurde. Auch beim Wiener ADS-Corpus (ADS steht für *adult-directed speech*) handelt es sich um eine sehr aktuelle Sprachdatensammlung, von etwas mehr als 57 Stunden Aufnahmen, an deren Transkription und Kodierung noch einige Zeit gearbeitet werden wird (Korecky-Kröll 2017).

3.8 *travel!digital*: Semantik mit anthroplogischem Blick

Im Rahmen des Projekts *travel!digital* wurde in den Jahren 2016–2017 eine digitale Sammlung früher deutschsprachiger Reiseführer außereuropäischer Länder (1875–1914) erstellt. Es entstand eine moderne in TEI-P5 ausgezeichnete Sprachressource, die neben linguistischen Basisannotationen (POS, Lemmata) auch semantische Informationen enthält. Inhaltlich wurde zu Fragen der kulturellen Repräsentation und identitätsstiftender Diskurse geforscht, wobei eine umfängliche Taxonomie der im Corpus erwähnten Personengruppen und Baudenkmäler erarbeitet wurde. Technologisch wurde mit zeitgemäßer RDF-basierter Technologie, insbesondere dem immer populärer werdenden Standard SKOS (*Simple Knowledge Organization System*) gearbeitet (Czeitschner & Krautgartner 2017). Als *go!digital*-Projekt war *travel!digital* von allem Anfang an als Teil der österreichischen CLARIN- und DARIAH-Aktivitäten konzipiert. Seit Dezember 2016 ist ein erstes Versuchsinterface online, das Teile dieses Corpus öffentlich zugänglich macht.[8]

---

8 https://baedeker.acdh.oeaw.ac.at/ (letzter Zugriff: 24. 10. 2017).

3.9 *Austrian Baroque Corpus* (ABaC:us)

Ein weiteres Beispiel für eine digitale Sprachressource, die nicht im engeren linguistischen Bereich entstanden ist, sehr wohl aber mit viel linguistischem Know-how erstellt und mit umfangreichen linguistischen Annotationen angereichert wurde, ist das *Austrian Baroque Corpus*, das mehrere Werke enthält, die dem Prediger und Schriftsteller Abraham a Sancta Clara (1644–1709) zugeschrieben werden (Resch & Czeitschner 2015). Die in ABaC:us vereinten historischen Texte stammen aus dem Zeitraum von 1650–1750 und lassen sich der Periode des älteren Neuhochdeutsch zuordnen. Das Corpus ist das Resultat interdisziplinärer Arbeit, in der historisch-literaturwissenschaftlich-linguistische Fragen mit texttechnologischen Interessen kombiniert wurden. Im Rahmen des Projekts *Texttechnologische Methoden zur Analyse österreichischer Barockliteratur*[9] wurde an Tools und Workflows für historische und nicht standardisierte linguistische Varietäten gearbeitet. Eines der Tools, die im Rahmen von ABaC:us intensiv weiterentwickelt wurden, ist der zuvor bereits beschriebene *tokenEditor*, mit dessen Hilfe die Gesamtheit der Wortklasseninformation des Corpus manuell überprüft werden konnte (Resch 2017). Die Textsammlung ist über eine Applikation online zugänglich und für wissenschaftliche Forschung frei nutzbar.[10]

4 Lexikographie

Auch immer mehr im akademischen Bereich kompilierte Wörterbücher werden unter Zuhilfenahme digitaler Hilfsmittel erzeugt und publiziert.

4.1 *Romani Lexicon Project* (ROMLEX)

Ein bemerkenswertes internationales Projekt, an dem österreichische Wissenschaftlerinnen und Wissenschaftler maßgeblich beteiligt waren und sind, ist das internationale *Romani Lexicon* Projekt (ROMLEX). ROMLEX ist eine lexikalische Datenbank, die eine Zahl von mindestens 25 Varietäten des Romani abdeckt. Die Einträge sind durchwegs mit englischen Übersetzungen resp. Übersetzungen in andere Sprachen (15 an der Zahl) versehen. Nach einer Auf-

---

9 Jubiläumsfonds der Österreichischen Nationalbank, Projekt Nr. 14738.
10 https://acdh.oeaw.ac.at/abacus/ (letzter Zugriff: 24. 10. 2017).

bauphase in den Jahren von 1998 bis 2001, in denen ausschließlich österreichische Varietäten aufgenommen wurden, wurde die Datenbank auf zahlreiche weitere Varietäten ausgeweitet (Halwachs, Schrammel & Rader 2006). Die Datenbank ist über ein online Interface abfragbar.[11]

4.2 TUNICO: die Schnittstelle zwischen Wörterbuch und Corpus

Seit mehreren Jahren arbeitet das Orientalistische Institut der Universität Wien in unterschiedlichen Projekten mit der ÖAW zusammen, zuerst mit dem Institut für Corpuslinguistik und Texttechnologie, später dann mit dem ACDH-OeAW, und war bestrebt, viele seiner Forschungen, auf eine moderne technologische Basis zu stellen. Eines dieser Projekte ist das *Vienna Corpus of Arabic Varieties* (VICAV), welches eine virtuelle auf die speziellen Bedürfnisse der Community zugeschnittene Forschungsplattform ist. Das Hauptinteresse liegt auf lexikalischen Informationen, die Beschreibung der linguistischen Varietäten erfolgt unter anderem durch standardisierte Sprachprofile, die neben Basisinformationen zu den jeweiligen Varietäten auch Eckdaten zur Forschungsgeschichte, zur verfügbaren Literatur etc. bieten. Neben einer umfassenden Bibliographie findet man auf der Webseite Wörterbücher, Glossare und unterschiedliche Arten transkribierter Texte (Procházka & Mörth 2017).

Ein Projekt, das aus VICAV heraus entwickelt wurde, ist TUNICO *(Linguistic dynamics in the Greater Tunis Area: a corpus-based approach)*. Methodologisch war das Projekt an der Schnittstelle zwischen traditioneller Dialektologie und moderner Sprachtechnologie angesiedelt und produzierte zwei digitale Sprachressourcen: ein Corpus des modernen gesprochenen Tunesischen und ein micro-diachrones Wörterbuch. Das Corpus, das sowohl Konversationen als auch Narrative enthält, ist das einzige umfangreichere transkribierte Corpus einer arabischen Varietät, das frei zugänglich ist.

Das TUNICO-Wörterbuch basiert zum Teil auf dem neuen Corpus, enthält aber auch lexikalisches Material, das aus anderen Quellen stammt. Es wurden sowohl gezielte Befragungen von Informanten als auch historische gedruckte Beschreibungen aus der Mitte und dem frühen 20. Jahrhundert in das Wörterbuch eingearbeitet. Ein spezieller Fokus des Projekts lag auf dem Interface zwischen Wörterbuch und Corpus, die in der technischen Implementierung eng verknüpft wurden. Es ist möglich, aus dem Wörterbuch heraus direkt ins Corpus zu navigieren, und umgekehrt sind die Wortformen des Corpus mit

11 http://romani.uni-graz.at/romlex/lex.xml (letzter Zugriff: 23.10.2017).

dem Wörterbuch verknüpft. Während des Projekts wurde viel Zeit in die manuelle Korrektur dieser Verlinkungen investiert (Mörth, Procházka & Dallaji 2014).

4.3 Deutschsprachige Lexikographie in Österreich

Die deutschsprachige Lexikographie ist in Österreich vor allem durch zwei ziemlich unterschiedliche lexikographische Unternehmungen repräsentiert, dem im akademischen Bereich angesiedelten *Wörterbuch der bairischen Mundarten in Österreich* (WBÖ) und dem *Österreichischen Wörterbuch* (ÖWB).[12]

Das Institut, welches an der Wiege des WBÖ stand, war die im Jahre 1911 gegründete *Kommission zur Schaffung des Österreichisch-Bayerischen Wörterbuches* an der Österreichischen Akademie der Wissenschaften. Inhaltlich steht das WBÖ in der Reihe der großlandschaftlichen Wörterbücher zur Erschließung der Varietäten des Deutschen. Während die Lieferungen dieses Wörterbuchs bislang nur in analoger Form erschienen, gab es seit 1993 gezielte Anstrengungen, die Materialbasis, den sogenannten Hauptkatalog des WBÖ, in digitaler Form zu erschließen. Hierzu wurde der Inhalt des Katalogs (geschätzte 2,5 Millionen Karteikarten) mit den historischen Belegen schrittweise transkribiert und in eine digitale Datenbank überführt.

Weder das Druckwerk noch die Datenbank als solche konnten bis dato zum Abschluss gebracht werden. Nachdem das Wörterbuch das Schicksal vieler vergleichbarer Projekte teilte und die Arbeiten Ende 2015 gänzlich zum Erliegen kamen, nahm die Österreichische Akademie der Wissenschaften im Frühjahr 2016 einen ganz neuen Anlauf und richtete am ACDH-OeAW eine eigene Forschungsabteilung unter dem Namen *Variation und Wandel des Deutschen in Österreich* ein, zu dessen Aufgaben die Fortsetzung des Wörterbuchprojekts zählt. Mit vollständig erneuerter Mannschaft wird das Projekt nunmehr unter Zuhilfenahme moderner Sprachtechnologie auf ganz neue Beine gestellt. Ziel ist es, lexikalische Daten sowohl analog als auch digital verfügbar zu machen. Das Projekt baut dabei auf die digitalen Infrastrukturen des ACDH-OeAW auf. Das Wörterbuch wird als Teil einer größeren webbasierten lexikographischen Forschungsplattform konzipiert.

---

12 Das Österreichische Wörterbuch (ÖWB) wird vom Österreichischen Bundesverlag (Ernst Klett-Verlag Leipzig/Stuttgart) produziert und vertrieben.

5 Terminologie

Die meisten der nachfolgenden terminologischen Ressourcen wurden zwar im akademischen Raum, zumeist im Rahmen von Forschungsprojekten, erstellt, haben aber sehr oft auch Nutzungsszenarien außerhalb der Forschung im Blick.

5.1 *AsylTermbank*

Im Rahmen des *AsylTerm*-Projektes, das von Oktober 2007 bis August 2008 vom Institut für Translationswissenschaft der Universität Graz und dem Zentrum für Translationswissenschaft der Universität Wien in Kooperation mit dem Bundesasylamt, dem Unabhängigen Bundesasylsenat und dem UN-Flüchtlingshochkommissariat UNHCR in Österreich durchgeführt wurde, wurde die Terminologie des österreichischen Asylverfahrens erfasst und eine Terminologiedatenbank zur österreichischen Asylterminologie, die sogenannte *AsylTermbank* angelegt, die online über ein Interface zugänglich ist (Termbase-Finder 2016). Gefördert wurde das Projekt durch den Europäischen Flüchtlingsfonds (EFF) und UNHCR Österreich. Das *AsylTerm*-Projekt hatte die Aufgabe, einen Beitrag zur Qualitätssicherung im Bereich translatorischer Dienstleistungen im Asylwesen zu leisten. Die *AsylTermbank* ist zwar eine mehrsprachige, aber monodirektionale Terminologiedatenbank. Ausgehend von der österreichischen Rechtsordnung wurde die österreichische Terminologie des Asylverfahrens deskriptiv erfasst und die Übersetzungen für Arabisch, Englisch, Französisch, Russisch und Serbisch festgelegt (Hebenstreit et al. 2009). Das heißt, die Terminologie in den anderen Sprachen ist nicht durch einen Rechtsvergleich auf Microebene zwischen dem österreichischen Asylrecht und den Zielrechtsordnungen entstanden (Chiocchetti et al. 2013)

5.2 *Innsbrucker Termbank 2.0*

Die *Innsbrucker Termbank 2.0* beinhaltet die terminologischen Diplom- und Masterarbeiten, die von Studierenden des Instituts für Translationswissenschaft der Universität Innsbruck verfasst wurden. Laut Webseite beinhaltet die Terminologiedatenbank zurzeit insgesamt 30.725 zwei- oder mehrsprachige Einträge aus 25 Fachbereichen in den Sprachen Deutsch, Englisch, Französisch, Italienisch, Spanisch und Russisch. In der Datenbank können Fachwörter zum Thema Abwasserreinigung und Baumaterialien genauso gefunden werden wie Termini aus dem Bereich der Käseproduktion, des Wertpapier-

handels oder des Sports. Da die Daten aus unterschiedlichen Diplom- und Masterarbeiten stammen, gibt es unterschiedliche Eintragsstrukturen. Jedoch beinhalten alle Einträge die Angabe des Fachbereichs, den ausgangssprachlichen und den zielsprachlichen Terminus sowie die Begriffsdefinition und die Quelle, aus der die Definition stammt (Innsbrucker Termbank 2.0 2015).

5.3 *RisikoTermbank*

Die *RisikoTermbank* entstand im Projekt *Wide Information Network to Improve Risk Management* (WIN). Das WIN-Projekt lief von 2004–2008, und ein Workpackage war auf die Erstellung von Tools zur Unterstützung der mehrsprachigen Kommunikation im Bereich Risiko- und Krisenmanagement ausgerichtet, an dem die Université Marc Bloch Strasbourg, die Universität Wien, die Universität Duisburg-Essen und Technische Universität Chemnitz beteiligt waren. Die *RisikoTermbank* hatte als Nutzergruppe Risikomanager und Bauingenieure im Blick, sollte aber auch von Lehrenden, Studierenden, Übersetzern und Übersetzerinnen sowie Journalistinnen und Journalisten genutzt werden. Die ca. 230 Begriffe aus dem Risikomanagement wurden aus einschlägiger Fachliteratur und aus einem mehrsprachigen Vergleichscorpus extrahiert (Budin 2011). Die Termbankeinträge sind über Begriffsbeziehungen miteinander verknüpft (z. B. *is cause of*, *has cause*). Die *RisikoTermbank* ist auf der Website des Zentrums für Translationswissenschaft öffentlich zugänglich (TermbaseFinder 2016).

6 Conclusio

Sowohl die Zahl an seriös produzierten digitalen Sprachdaten als auch an digital gestützten Projekten ist in den letzten Jahren stark gestiegen. In vielen Bereichen der Forschung ist der Einsatz digitaler Sprachressourcen nicht mehr wegzudenken und mittlerweile Teil des Standardrepertoires an zeitgemäßen Forschungsmethoden geworden. Positiv zu vermerken ist des Weiteren, dass das Bewusstsein um *Open Access* und *Open Source* immer verbreiteter wird und dadurch immer mehr nachnutzbare Ressourcen verfügbar werden.

Literatur

Atz, Stefan & Anita Elisabeth Gerstbauer (2017): *Emerging markets in the Austrian press an exploratory study at the intersection of international business research and computer-based language analysis.* Masterarbeit. Wirtschaftsuniversität Wien.

Bellini, Daniele & Stefan Schneider (2006): Spoken Italian online: The *Banca dati dell'italiano parlato* (BADIP). In Bernhard Kettemann, Georg Marko (Hrsg.), *Planning, gluing and painting corpora. Inside the applied corpus linguist's workshop,* 13–26. Frankfurt a. M.: Lang.

Blumesberger, Susanne (2010): Phaidra. Digitale Langzeitarchivierung an der Universität Wien. In Christiane Fennesz-Juhasz et al. (Hrsg.), *Digitale Verfügbarkeit von audiovisuellen Archiven im Internet-Zeitalter. Beiträge zur Tagung der Medien Archive Austria und des Phonogrammarchivs der ÖAW. Dietrich Schüller zum 70. Geburtstag,* 77–84. Berlin: LIT.

Breiteneder, Angelika, Theresa Klimpfinger, Stefan Majewski & Marie-Luise Pitzl (2009): The Vienna-Oxford International Corpus of English (VOICE). A linguistic resource for exploring English as a lingua franca. *ÖGAI Journal* 28 (1): 21–26.

Budin, Gerhard (2011): Designing and implementing strategies of global, multilingual "Disaster Communication". In Larissa Alekseeva (Hrsg.), *Proceedings of the 18th European Symposium on Language for Special Purposes (LSP)*, 11–26. Perm: Perm State University.

Czeitschner, Ulrike & Barbara Krautgartner (2017): Discursive constructions of culture: Semantic modelling for historical travel guides *Sociology and Anthropology*. 5 (4): 323–331. doi: 10.13189/sa.2017.050406. http://www.hrpub.org (letzter Zugriff: 24. 10. 2017).

Chiocchetti, Elena, Vesna Lusicky, Natascia Ralli & Tanja Wissik (2013): Spanning bridges between theory and practice. Terminology workflow in the legal and administrative domain. *Comparative Legilinguistics. International Journal of Legal Communication*. 16, 7–22.

GAMS (2017): Dokumentation. In: *GAMS Geisteswissenschaftliches Asset Management System*. http://gams.uni-graz.at/docs (letzter Zugriff: 25. 4. 2017).

Halwachs, Dieter W., Barbara Schrammel & Astrid Rader (2006): ROMLEX – the Lexical Database of Romani Varieties. http://glm.uni-graz.at/etc/publications/GRP-Halwachs-Schrammel-Rader-2006.pdf (letzter Zugriff: 24. 10. 2017).

Hebenstreit, Gernot, Sonja Pöllabauer & Irmgard Soukup-Unterweger (2009): AsylTerm: Terminologie für Dolmetscheinsätze im Asylverfahren. *trans-kom* 2 (2), 173–196.

Innsbrucker Termbank 2.0 (2015): https://orawww.uibk.ac.at/apex/uprod/f?p=201206111:1:0::NO::: (letzter Zugriff: 25. 4. 2017).

Korecky-Kröll, Katharina (2017): Kodierung und Analyse mit CHILDES: Erfahrungen mit kindersprachlichen Spontansprachkorpora und erste Arbeiten zu einem rein erwachsenensprachlichen Spontansprachkorpus. In Claudia Resch & Wolfgang U. Dressler (Hrsg.), *Digitale Methoden der Korpusforschung in Österreich*, 85–113. Wien: Verlag der ÖAW (Sitzungsberichte der phil.-hist. Klasse. Band 897).

Lusicky, Vesna & Tanja Wissik (2017): Discovering resources in the VLO: A pilot study with students from translation studies. In *Selected papers from the CLARIN Annual Conference 2016*, 63–75. Linköping: Linköping University Electronic Press.

Mörth, Karlheinz, Stephan Procházka & Ines Dallaji (2014): Laying the foundations for a diachronic dictionary of Tunis Arabic. A first glance at an evolving new language resource. In A. Abel, C. Vettori & N. Ralli (Hrsg.), *Proceedings of the XVI EURALEX International Congress: The User in Focus*, 377–387. Bolzano: EURAC research.

Osimk-Teasdale, Ruth (2013): Applying existing tagging practices to VOICE. In Mukherjee Joybrato & Magnus Huber (Hrsg.), *Corpus linguistics and variation in English: Focus on nonnative Englishes (Proceedings of ICAME 31)*, Helsinki: VARIENG. http://www.helsinki.fi/varieng/series/volumes/13/osimk-teasdale/ (letzter Zugriff: 24. 10. 2017).

Pessentheiner, Hannes, Thomas Pichler & Martin Hagmüller (2016): AMISCO: The Austrian German Multi-Sensor Corpus. In N. Calzolari et al. (Hrsg.), *Proceedings of*

the 10th International Conference on Language Resources and Evaluation (LREC'16), 760–766. Portorož, Slovenia: ELRA.

Procházka, Stephan & Karlheinz Mörth (2017): The Vienna Corpus of Arabic Varieties: building a digital research environment for Arabic dialects. In M. Al-Hamad, R. Ahmed & H. Aloui (Hrsg.), *Lisan Al-Arab: Studies in contemporary Arabic dialects, Proceedings of the 10th International Conference of AIDA*, *Qatar University 2013*, 176–183. Wien: LIT. (Forthcoming)

Ransmayr, Jutta, Karlheinz Mörth & Matej Ďurčo (2013): Linguistic variation in the Austrian Media Corpus. Dealing with the challenges of large amounts of data. In Chelo Vargas-Sierra (Hrsg.), *Corpus resources for descriptive and applied studies. Current challenges and future directions: Selected papers from the 5th International Conference on Corpus Linguistics (CILC2013),* 111–115. Oxford: Elsevier (Procedia – Social and Behavioral Sciences 95).

Ransmayr, Jutta, Sonja Schwaiger, Matej Ďurčo, Hannes Pirker & Wolfgang U. Dressler (2016): Graduierung der Transparenz von Diminutiven auf -chen: Eine korpuslinguistische Untersuchung. *Deutsche Sprache. Zeitschrift für Theorie, Praxis, Dokumentation* 44/3: 261–286.

Resch, Claudia & Ulrike Czeitschner (Hrsg.) (2015): ABaC:us – Austrian Baroque Corpus. http://acdh.oeaw.ac.at/abacus (letzter Zugriff: 25. 4. 2017).

Resch, Claudia (2017): „Etwas für alle" – Ausgewählte Texte von und mit Abraham a Sancta Clara digital. *Zeitschrift für digitale Geisteswissenschaften*. doi: 10.17175/2016_005.

Schuppler, Barbara, Martin Hagmueller, Juan A. Morales-Cordovilla & Hannes Pessentheiner (2014): GRASS: the Graz Corpus of Read and Spontaneous Speech. In N. Calzolari et al. (Hrsg.), *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, 1465–1470. Reykjavik: ELRA.

Simons, Gary F. & Steven Bird (2008): Toward a Global Infrastructure for the Sustainability of Language Resources. In *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation*, 87–100. Cebu City, Philippines: De La Salle University.

TermbaseFinder (2016): *TermbaseFinder am Zentrum für Translationswissenschaften der Universität Wien*. http://termbasefinder.trans.univie.ac.at (letzter Zugriff: 25. 4. 2017).

Tošović, Branko (2017): Die morphologische Annotation im Gralis-Korpus. In Claudia Resch & Wolfgang U. Dressler (Hrsg.), *Digitale Methoden der Korpusforschung in Österreich*, 63–83. Wien: Verlag der ÖAW. (Sitzungsberichte der phil.-hist. Klasse. Band 897).

Transbank (2017): *Projektwebseite des Transbank Projekts*. https://transbank.info (letzter Zugriff: 25. 4. 2017).

Transkribus (2016): *Wiki*. https://transkribus.eu/wikiDe/index.php/Hauptseite (letzter Zugriff: 25. 4. 2017).

Trippel, Thorsten, Thierry Declerck & Ulrich Heid (2015): Sprachressourcen in der Standardisierung. *Journal for Computational Linguistics and Language Technology* 20 (2): 17–30.

Werner, Martina (2017): Zur Entwicklung der synthetischen Komposition in der Geschichte des Deutschen. In J. Meibauer et al. (Hrsg.), *Special Issue on Zusammenbildungen/ Synthetic compounds. Zeitschrift für Wortbildung/Journal of Word-Formation* 1: 73–92.

Wissik, Tanja & Gerhard Budin (2010): *CLARIN-AT – Project Report. Erhebung Sprachressourcen und Sprachtechnologien in Österreich*, Juli 2010. Wien: Universität Wien.

Wynne, Martin (2015): *User Involvement. Vortrag auf der Clarin Annual Conference 2015.* https://www.clarin.eu/sites/default/files/20151016-CAC-04-Wynne-User-Involvement-CAC2015-05.pdf (letzter Zugriff: 25. 4. 2017).



II Sprachwissenschaft und Sprachtechnologie

Hannah Kermes und Elke Teich

5 Generische Infrastruktur und spezifische Forschung: Angebote und Lösungen

Abstract: Die empirische Forschung an natürlichsprachlichen Daten geht mit grundlegenden methodischen Veränderungen einher. Immer mehr Texte stehen in digitaler Form zu Verfügung. Eine rein manuelle Vorgehensweise ist nicht möglich oder extrem zeitaufwendig. Wir zeigen welche Vorteile der Einsatz von generischen Infrastrukturkomponenten für spezifische Forschung haben kann: (i) effiziente Untersuchungen auf größeren Datenmengen, (ii) reproduzierbare und übertragbare Ergebnisse. Wir zeigen an einer konkreten Studie, wie generische Infrastruktur spezifisch angepasst und durch spezifische Lösungen ergänzt werden kann.

Keywords: Annotation, Empirie, Sprachkorpora, Textanalyse, XML

1 Einleitung

Die empirische Forschung an natürlichsprachlichen Daten geht mit grundlegenden methodischen Veränderungen einher (cf. Biber, Conrad & Reppen 1998: Kapitel 1; McEnery, Xiao & Tono 2006: 3–4). Immer mehr Texte stehen in digitaler Form zu Verfügung, ob über Plattformen wie Wikisource,[1] durch Projekte wie Project Gutenberg,[2] das Deutsche Text Archiv[3] oder das Linguistic Data Consortium.[4] Es gilt, im Hinblick auf eine Forschungsfrage in einer großen Menge an Ausgangsdaten Relevantes zu finden und zu extrahieren. Das Resultat sind oft große, zumeist multidimensionale Datensätze, die analysiert

1 http://wikisource.org/ (letzter Zugriff: 22. 4. 2018).
2 http://www.gutenberg.org/ (letzter Zugriff: 22. 4. 2018).
3 http://www.deutschestextarchiv.de/ (letzter Zugriff: 22. 4. 2018).
4 https://www.ldc.upenn.edu/ (letzter Zugriff: 22. 4. 2018).

Anmerkung: Die im Artikel beschriebenen Arbeiten wurden durch das Bundesministeriums für Bildung und Forschung im Rahmen des CLARIN-D Projekts unterstützt. Besonderer Dank gilt unseren Kollegen Peter Fankhauser, Stefan Fischer und Jörg Knappen.

Hannah Kermes, Universität des Saarlandes, Fachrichtung Sprachwissenschaft und Sprachtechnologie, Campus A2.2, D-66123 Saarbrücken, E-Mail: h.kermes@mx.uni-saarland.de
Elke Teich, Universität des Saarlandes, Fachrichtung Sprachwissenschaft und Sprachtechnologie, Campus A2.2, D-66123 Saarbrücken, E-Mail: e.teich@mx.uni-saarland.de

https://doi.org/10.1515/9783110538663-006

und schließlich interpretiert werden müssen. Eine rein manuelle Vorgehensweise ist in der Regel nicht möglich oder extrem zeitaufwendig. Zudem macht die Komplexität der einzelnen Verarbeitungsschritte den Einsatz von generischen bzw. automatischen Werkzeugen notwendig (cf. McEnery, Xiao & Tono 2006: Kapitel 1; Lemnitzer & Zinsmeister 2010: Kapitel 4). Nicht zuletzt ist der Einsatz von generischer Infrastruktur auch bezüglich Reproduzierbarkeit und Übertragbarkeit sinnvoll.

Forschungsinfrastrukturen wie CLARIN-D[5] und Plattformen wie Gate[6] oder NLTK[7] haben sich darauf spezialisiert, sprachtechnologische Werkzeuge und Sprachressourcen zur linguistischen Annotation (z. B. Wortartenannotation, syntaktischen oder semantischen Annotation) und zur Textanalyse für eine breite Zielgruppe zugänglich zu machen. Die Problematik liegt darin, die Angebote generischer Infrastruktur für die speziellen Anforderungen der eigenen Forschung nutzbar zu machen. Es gilt daher zunächst generische und spezifische Komponenten einer Studie zu identifizieren. Generische Komponenten sind z. B. vorhandene Corpora, Corpusabfragewerkzeuge oder Werkzeuge für die Corpusanalyse, aber auch Werkzeuge für die linguistische Annotation oder die OCR-Fehlerkorrektur. Spezifische Komponenten sind etwa das zu untersuchende linguistische Phänomen, aber auch eine spezielle Textgrundlage. Dabei stellen sich folgende Fragen: Wo können generische Komponenten eingesetzt werden, wo müssen spezifische Lösungen gefunden werden und wie können generische und spezifische Komponenten miteinander verknüpft werden?

Im Folgenden werden wir zunächst die allgemeine Vorgehensweise bei einer corpuslinguistischen Studie im Hinblick auf das Zusammenspiel von generischer Infrastruktur und spezifischen Lösungen diskutieren. Danach werden wir anhand einer konkreten Studie beschreiben, wie dieses Zusammenspiel in der Realität aussehen kann, von der Aufbereitung der Ausgangsdaten (Vorverarbeitung und Annotation des Corpus (Abschnitt 3.1) bis zur Datenanalyse (Abschnitt 3.4).

2 Methodik, Arbeitsabläufe und Angebote

In der Corpuslinguistik wird ein sprachliches Phänomen quantitativ und qualitativ anhand einer geeigneten Textgrundlage untersucht. Dazu werden rele-

5 https://www.clarin-d.de/ (letzter Zugriff: 22. 4. 2018).
6 https://gate.ac.uk/ (letzter Zugriff: 22. 4. 2018).
7 http://www.nltk.org (letzter Zugriff: 22. 4. 2018).

vante Beobachtungen extrahiert und anschließend analysiert und interpretiert. Die Auswahl der Textgrundlage ist dabei von dem zu untersuchenden Phänomen abhängig. Existiert kein geeignetes Corpus, so muss eines erstellt werden. Corpuslinguistische Studien lassen sich daher in zwei Hauptbereiche aufteilen: (i) die Corpuserstellung und (ii) die Corpusanalyse. Im Folgenden wollen wir nun darauf eingehen, wie generische Infrastruktur diese beiden Bereiche unterstützen kann und wo spezifische Lösungen gefunden werden müssen.

2.1 Corpuserstellung

Die Corpuserstellung gliedert sich in zwei Schritte: (i) die Vorverarbeitung und (ii) die linguistische Annotation. Dabei gehen wir von bereits digitalisierten Texten[8] aus und nehmen auch an, dass die Texte bereits ausgewählt und zu einer Textsammlung (Corpus) zusammengestellt wurden (cf. Lemnitzer & Zinsmeister 2010: Kapitel 3; McEnery, Xiao & Tono 2006: Kapitel 2). Eine solche digitalisierte aber ansonsten noch nicht weiterverarbeitete Textsammlung bezeichnen wir als Ausgangsdaten.

Bei der Weiterverarbeitung sind nun zwei Aspekte zu berücksichtigen. Erstens sind in der Regel in neu zusammengestellten Corpora Texte anderer Textsorten, Register, Zeitperioden oder Autoren enthalten als in bereits existierenden Corpora. Die über eine Infrastruktur bereitgestellten Verarbeitungskomponenten, wie z. B. Wortartentagger, sind aber in der Regel auf der Allgemeinsprache (z. B. auf Zeitungstexten) trainiert. Je mehr die Texte in einem neu zusammengestellten Corpus von der Allgemeinsprache abweichen, desto wahrscheinlicher ist es, dass generische Komponenten nicht das gewünschte Ergebnis liefern (geringe Abdeckung, hohe Fehlerrate). In diesem Fall müssen die generischen Komponenten spezifisch angepasst bzw. ergänzt werden.

Im Folgenden werden die einzelnen Komponenten der Corpuserstellung anhand der schematischen Darstellung in den Abbildungen 5.1 (Vorverarbeitung) und 5.2 (linguistische Annotation) hinsichtlich des Einsatzes von generischen Komponenten und spezifischen Lösungen näher diskutiert, wobei sowohl auf automatische Werkzeuge als auch auf manuelle Methoden eingegangen wird. Dabei gehen wir zunächst auf die Vorverarbeitung ein (Abschnitt 3.2), die sich in Datenkonversion (Überführung der Ausgangsdaten in ein standardisiertes Format) und Datenbereinigung (Lösung von Layoutproblemen und OCR Fehlerkorrektur) gliedert. Auf die Anreicherung mit Metadaten,

---

8 Als ein Beispiel für den Arbeitsablauf bei der Digitalisierung von Texten sei auf den Digitalisierungsworkflow des Deutschen Text Archivs verwiesen, http://www.deutschestextarchiv.de/doku/workflow (letzter Zugriff: 22. 4. 2018).

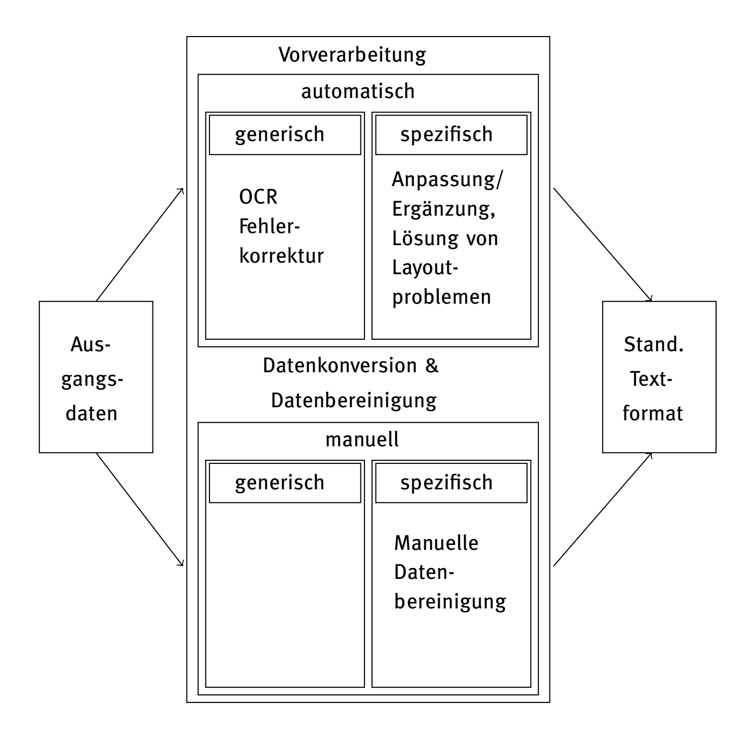


Abb. 5.1: Arbeitsablauf bei der Vorverarbeitung.

also mit das Corpus beschreibende Daten wie Autor, Titel, Erscheinungsjahr etc., die auch zur Datenkonversion gehört, wird hier nicht näher eingegangen. In Abschnitt 3.3 wird dann die linguistische Annotation beschrieben, die sich wiederum in verschiedene Annotationsebenen (Wortebene, syntaktische Ebene) und automatische und manuelle Methoden der Annotation gliedert.

2.1.1 Vorverarbeitung

Die Ausgangsdaten eines Corpus liegen nur selten in einem standardisierten Format vor, z. B. in einem TEI-konformen (*Text Encoding Initiative*[9]) XMLFormat oder einem einfachen Textdokument mit optionalem TEI-konformem XML-

9 "The Text Encoding Initiative (TEI) Guidelines are an international and interdisciplinary standard that facilitates libraries, museums, publishers, and individual scholars represent a variety of literary and linguistic texts for online research, teaching, and preservation." http://www.tei-c.org/ (letzter Zugriff: 22. 4. 2018): siehe u. a. die Kapitel 5 (*TEI Header*) und 23 (*Language Corpora*).

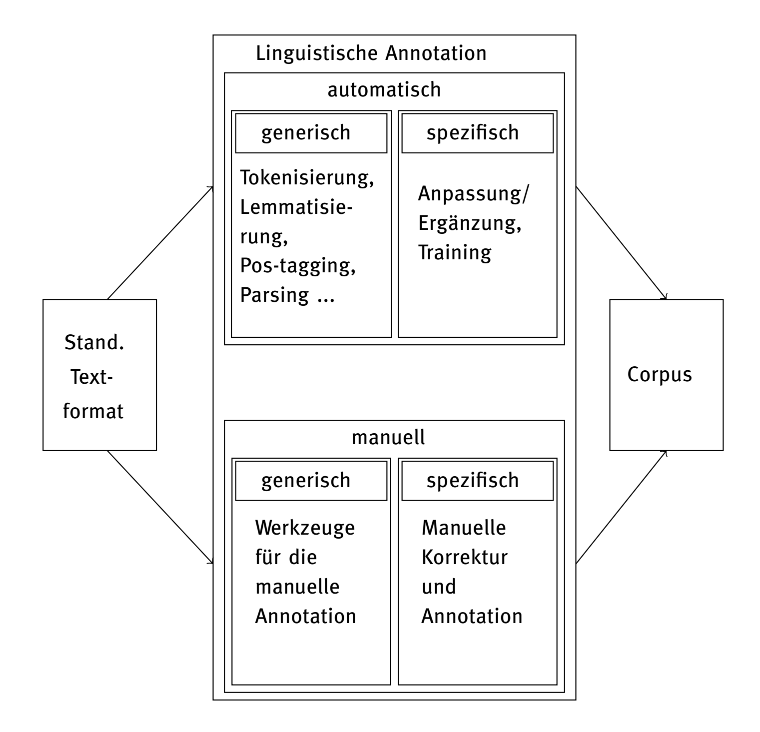


Abb. 5.2: Arbeitsablauf bei der linguistischen Annotation.

Markup. Ein wichtiger Schritt bei der Vorverarbeitung ist die Datenkonversion, also die Überführung der spezifischen Ausgangsdaten in ein standardisiertes (generisches) Format. Standardisierte Daten sind für die Weiterverarbeitung und die Wiederverwendung entscheidend.

Standardisierung bedeutet aber nicht, dass die Formate nicht auf spezifische Bedürfnisse angepasst werden können. Je nach intendiertem Verwendungszweck kann das Format daher variieren und spezifische Kriterien aufweisen.

Ein Beispiel für ein solche Spezifikation auf der Basis eines TEI Formats ist das DTA-Basisformat,[10] dessen Richtlinien auf der einen Seite eine umfassende Textaufbereitung erlauben und auf der anderen Seite die Flexibilität bei der Annotation so einschränken, dass die entstehenden Texte insgesamt kohärent sind. Es ist auf die Annotation gedruckter historischer Texte spezialisiert und dient als Basis sowohl für digitale Editionen als auch für Textcorpora.

10 http://www.deutschestextarchiv.de/doku/basisformat (letzter Zugriff: 22. 4. 2018).

Ein weiteres Beispiel ist das TCF Format,[11] ein XML Format für Textcorpora mit multidimensionaler linguistischer Annotation, das als Austauschformat für WebLicht[12] (cf. Kapitel 3.3), einer Plattform für die automatische Annotation von Textcorpora, entwickelt wurde.

Das CONLL-Format (Buchholz & Marsi 2006) und das VRT-Format der IMS Corpus Workbench (CWB)[13] sind Beispiele für textbasierte Standardformate (*one-word-per-line*). Annotationen sind entweder als TAB-getrennte Spalten kodiert oder beim VRT-Format auch zusätzlich als XML-Elemente. Durch ihren einerseits standardisierten und andererseits einfachen und flexiblen Aufbau sind sie sehr gut als Austauschformate geeignet.

Ein weiterer Aspekt der Vorverarbeitung ist die Datenbereinigung, die sich um Fehlerkorrektur und das Entfernen nicht erwünschter Elemente (Rauschen) kümmert. Die Datenbereinigung hat sowohl generische als auch spezifische Aspekte. Zu den spezifischen Aspekten gehören Layoutprobleme wie Kopf- und Fußzeilen, Graphiken und Tabellen. Obwohl diese Probleme bei vielen Corpora auftreten, sind ihre Ausprägungen und damit auch die Erkennungsmuster sehr unterschiedlich. So muss hier auf spezifische Lösungen zurückgegriffen werden, etwa auf dedizierte Skripte oder auch manuelle Korrektur bei der Entfernung von nicht-textuellen Elementen wie Graphiken oder Tabellen.

Fehler, die aufgrund einer automatischen Texterkennung (OCR-Fehler) entstanden sind, können sowohl generisch als auch spezifisch sein. Für typische OCR-Fehler gibt es eine Reihe von generischen Komponenten zur Korrektur, wobei es sich im Wesentlichen um Ersetzungslisten handelt. Ein Beispiel ist die Ersetzungsliste von Underwood & Auvil (2012) mit 50.000 Ersetzungspaaren für historische englische Texte aus den Jahren 1700–1899. Ein weiteres Beispiel ist das Projekt OCR-D,[14] das auf deutsche historische Texte spezialisiert ist. Das Projekt hat eine ganzheitliche Lösung zum Ziel und setzt bereits bei der OCR-Erkennung selbst an. Für die OCR-Korrektur soll eine Datenbank mit Ersetzungsregeln und Ersetzungsmustern aufgebaut werden, die eine möglichst breite Abdeckung hat. Die generischen Komponenten können jedoch nur die typischen OCR-Fehler abdecken. Insbesondere bei sehr spezifischen Corpora (z. B. Spezialvokabular, spezielle Renderings) kann eine Ergänzung oder Anpassung der Regeln notwendig sein, um zufriedenstellende Ergebnisse zu erzielen. Für die Evaluierung und eine etwaige Identifizierung von

11 https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format (letzter Zugriff: 22. 4. 2018).

12 http://weblicht.sfs.uni-tuebingen.de/ (letzter Zugriff: 22. 4. 2018).

13 http://cwb.sourceforge.net (letzter Zugriff: 22. 4. 2018).

14 http://www.ocr-d.de/ (letzter Zugriff: 22. 4. 2018).

spezifischen OCR-Fehlern können wiederum generische Komponenten aus dem Language Modeling, wie etwa *Word Embeddings*, eingesetzt werden (cf. Knappen et al. 2017).

Das konvertierte und bereinigte Corpus kann dann als Basis für die Weiterverarbeitung eingesetzt werden, also z. B. für die linguistische Annotation des Corpus.

2.1.2 Linguistische Annotation

Die linguistische Annotation fügt dem Corpus weitere Abstraktionsebenen hinzu, die auf einer Interpretation der Wörter oder Wortsequenzen in ihrem Kontext basieren. So können einerseits Abfragen effizienter gestaltet werden, weil sie präziser formuliert werden können und andererseits komplexe Phänomene automatisch extrahiert werden (cf. Lemnitzer & Zinsmeister 2010: 60–62). Die unterste Ebene der Annotation ist in der Regel die Annotation auf der Wortebene, auch morphosyntaktische Annotation genannt, und besteht aus Tokenisierung, Lemmatisierung und Wortartentagging (Schmid 2008, Voutilainen 2003; Leech & Wilson 1996). Auf ihr bauen die anderen Annotationsebenen auf, wie etwa die syntaktische Annotation in Form von (partiellen) Dependenz- oder Konstituentenstrukturannotationen (Manning & Schütze 1999: Chapter 12; Langer 2001; Kermes 2008), die semantische Annotation (z. B. Named Entity Recognition, semantische Rahmen, Lesarten) und die pragmatische Annotation (Anaphernund Koreferenzauflösung, Annotation von Informationsstruktur) (cf. Lemnitzer & Zinsmeister 2010: 84–86).

Die Annotation von linguistischer Information ist in jedem Fall aufwendig. Der Einsatz von automatischen Werkzeugen ist daher sinnvoll. Für die verschiedenen Annotationsebenen gibt es eine ganze Reihe von generischen Komponenten (cf. Lemnitzer & Zinsmeister 2010; McEnery, Xiao & Tono 2006; Kübler & Zinsmeister 2015). Forschungsinfrastrukturen wie CLARIN-D stellen eine Auswahl dieser Werkzeuge webbasiert zur Verfügung mit der Möglichkeit eigene Prozessketten zusammenzustellen (cf. WebLicht; Hinrichs, Hinrichs & Zastrow 2010, Düsendi 2014). Plattformen wie Gate[15] (*General architecture for text engineering*, Cunningham, Maynard & Bontcheva 2011; Cunningham et al. 2013) oder NLTK (*Natural Language Tool Kit*)[16] stellen ähnliche Funktionalitäten zur Verfügung.

15 https://gate.ac.uk/ (letzter Zugriff: 22. 4. 2018).
16 http://www.nltk.org/ (letzter Zugriff: 22. 4. 2018).

Beim Einsatz dieser generischen Komponenten ist zu beachten, dass die Annotationen, die automatische Werkzeuge liefern nicht perfekt sind (d. h. es treten Fehler auf), da die Werkzeuge für eine bestimmte Sprachvarietät optimiert sind, zumeist für Allgemeinsprache (s. oben). Eine Evaluierung der Annotation kann die Qualität der Annotation bestimmen und Probleme aufzeigen. Die Qualität der Annotation bemisst sich nach dem Anteil der korrekten Annotation an der Anzahl aller Annotationen (der sogenannten Präzision). Ist die Qualität der Annotation zu schlecht, muss das generische Werkzeug eventuell angepasst oder neu trainiert werden. In der Regel bedeutet dies aber auch, dass eine webbasierte Prozesskette nicht mehr möglich ist, da hier nur die generischen Komponenten verwendet werden können. Die Lösung ist in diesem Fall der Aufbau einer eigenen lokalen Prozesskette mit angepassten generischen Komponenten.

Bei kleineren Corpora besteht auch die Möglichkeit manuell zu annotieren oder die Annotation manuell zu korrigieren. Auch hier gibt es generische Werkzeuge, die die manuelle Annotation unterstützen, wie etwa MMAX (Müller & Strube 2001), annotate (Brants & Plaehn 2000), SALTO für semantische Annotationen (Burchardt et al. 2006) oder EXMARaLDA[17] für Sprachcorpora. Einige davon, etwa das webbasierte Werkzeuge WebAnno,[18] unterstützen eine Vielzahl von Annotationsebenen, darunter auch selbst definierte Ebenen. WebAnno hat zudem einen automatischen Modus, der lernt und Annotationen vorschlägt (cf. Eckart de Castilho et al. 2014; Yimam et al. 2013, 2014). Die Entscheidung wieviel im Einzelfall in spezifische Lösungen investiert wird, ist immer eine Abwägung zwischen Qualität und Effizienz.

2.2 Corpusanalyse

Die spezifischste Komponente einer corpuslinguistischen Studie ist natürlich die Analyse der extrahierten Daten (z. B. eine Merkmalsverteilung). Aber auch bei der Corpusanalyse lassen sich nicht nur spezifische sondern auch generische Elemente identifizieren. Abbildung 5.3 zeigt eine schematische Darstellung des Arbeitsablaufes bei der Corpusanalyse.

Prinzipiell kann man zwei Vorgehensweisen unterscheiden, *corpus-based* (corpusbasiert oder phänomengesteuert) und *corpus-driven* (corpusgesteuert oder explorativ) (cf. Tognini-Bonelli 2001: 30; McEnery, Xiao & Tono 2006: 8–10). Bei der corpusbasierten Vorgehensweise werden zunächst die für die Un-

17 http://exmaralda.org/de/ (letzter Zugriff: 22. 4. 2018).
18 https://webanno.github.io/webanno/ (letzter Zugriff: 22. 4. 2018).

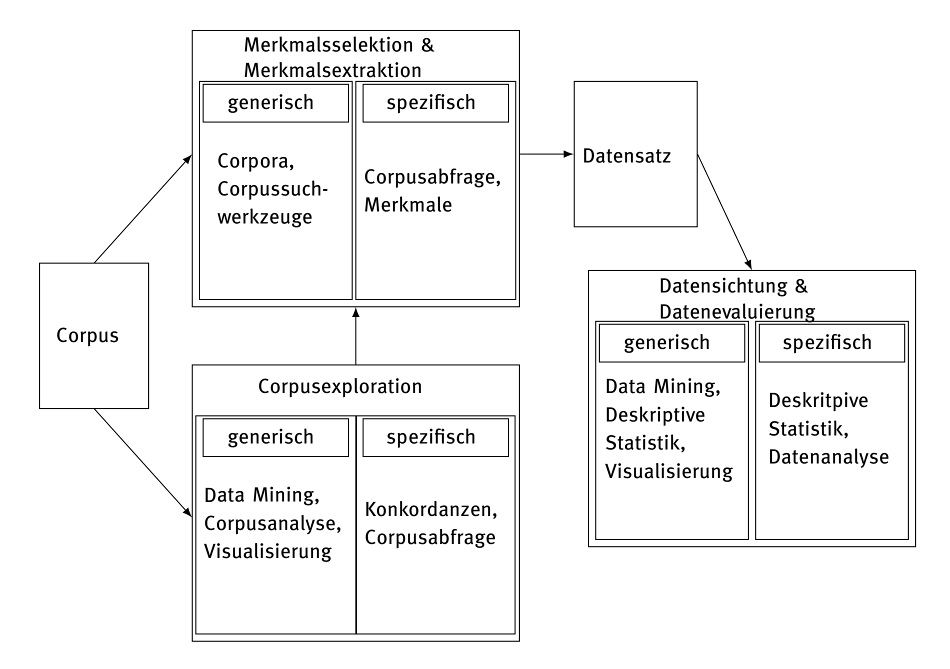


Abb. 5.3: Arbeitsablauf bei der Corpusanalyse.

tersuchung relevanten Merkmale identifiziert (Merkmalssselektion) (cf. Biber & Finegan 2014) und anschließend die entsprechenden Corpusinstanzen extrahiert (Merkmalsextraktion). Der daraus resultierende Datensatz aus Feature-Wert-Paaren wird dann gesichtet (z. B. mit Hilfe von Visualisierungen) und evaluiert (z. B. mit Methoden der deskriptiven Statistik oder Data Mining) und schließlich interpretiert.

Die corpusgesteuerten Vorgehensweise basiert auf einer explorativen Analyse des Corpus z. B. mit dem Ziel typische Merkmale für eine durch das Corpus repräsentierte Sprachvarietät zu identifizieren. Generische Komponenten der Corpusexploration beinhalten oft die einzelnen Schritte von der Merkmalsselektion, über die Merkmalsextraktion bis zur Datensichtung und Datenevaluierung. Die Corpusexploration kann aber auch als Unterstützung bei der Merkmalsselektion der corpusbasierten Methode dienen (cf. Biber & Finegan 2014). Beide Methoden ergänzen einander (McEnery, Xiao & Tono 2006).

Bei der Corpusexploration können generische Werkzeuge, z. B. aus dem Language Modeling (Word Embeddings, N-gram-Modelle) oder Corpusanalysetools wie das Voyant Tool[19] eine andere Sichtweise auf das Corpus ermög-

19 http://voyant-tools.org/ (letzter Zugriff: 22. 4. 2018).

lichen. Sie abstrahieren von der Textgrundlage durch (i) Gruppierungen in Form von Häufigkeitsverteilungen von Wörtern und Phrasen, Messungen der lexikalischen Dichte und Varianz und (ii) Hervorhebung von typischen Elementen in Form von Termen, Keywords oder grammatikalischen Einheiten. Aber auch die manuelle Sichtung eines Corpus über Corpussuche und Konkordanzen kann vor allem bei kleineren Corpora sinnvoll sein. Durch die Corpusexploration können auch noch unbekannte relevante Merkmale identifiziert werden. Für die detaillierte Analyse der Merkmale muss dann eine Datensichtung auf einer Mikroebene stattfinden, etwa durch Extraktion und/oder Datensichtung und Datenevaluierung auch auf der Textebene (Konkordanzen).

Bei der corpusbasierten Methode geht man von einem zu untersuchenden linguistischen Phänomen aus. Die Merkmalsselektion erfolgt manuell auf der Basis von vorhandenen linguistischen Studien (Merkmalskatalog als generische Ressource) aber auch über die Corpusexploration. Für die Merkmalsextraktion werden in der Regel generische Corpusabfragewerkzeuge verwendet, die Abfragen selbst sind natürlich spezifisch, aber es kann sich lohnen wiederkehrende Abfragen in einer Art Abfragebibliothek zu speichern. Inzwischen sind viele Corpora online verfügbar und abfragbar. Dabei unterscheiden sich die einzelnen Plattformen für die Corpusabfrage hinsichtlich ihrer Funktionalität. Exemplarisch seien hier für die deutsche Sprache COSMAS (*Corpus Search, Management and Analysis System*[20]) als Abfragewerkzeug für die Corpussammlung des Instituts für Deutsche Sprache, sowie das DWDS (*Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart*[21]) und das Deutsche Text Archiv[22] genannt. Für englischsprachige Corpora seien die BYU Corpora[23] sowie das BNCweb[24] und die Corpora auf der CQPweb Plattform der Lancaster University[25] genannt. Schließlich sei noch das OPUS Projekt (*open parallel corpus*[26]) erwähnt, das eine große Sammlung von parallelen Übersetzungstexten als alignierte Corpora bereitstellt.

Die meisten der bereitgestellten Corpora sind lemmatisiert und wortartengetaggt. Corpusabfragen können auf diesen Annotationen aufsetzen und erlauben den Einsatz von regulären Ausdrücken bei der Abfrage. Je nach Untersuchungsgegenstand bieten diese online Plattformen unterschiedliche

20 http://www.ids-mannheim.de/cosmas2/ (letzter Zugriff: 22. 4. 2018).
21 https://www.dwds.de/ (letzter Zugriff: 22. 4. 2018).
22 http://www.deutschestextarchiv.de/ (letzter Zugriff: 22. 4. 2018).
23 http://corpus.byu.edu/ (letzter Zugriff: 22. 4. 2018).
24 http://corpora.lancs.ac.uk/BNCweb/ (letzter Zugriff: 22. 4. 2018).
25 https://cqpweb.lancs.ac.uk/ (letzter Zugriff: 22. 4. 2018).
26 http://opus.lingfil.uu.se/ (letzter Zugriff: 22. 4. 2018).

generische Lösungen für die Corpusanalyse. Als Ergebnis wird immer zumindest eine Konkordanz ausgegeben und die Anzahl der Treffer. Bei OPUS können zusätzlich noch alignierte Textstellen mit ausgegeben werden.

Einige der Plattformen bieten darüber hinaus weitere Auswertungsmöglichkeiten. Das DWDS bietet für Einzelwörter Zugriff auf lexikalische Informationen wie Bedeutung, Etymologie, Thesaurus, typische Verbindungen sowie eine Verlaufskurve der Häufigkeit des Wortes über die Zeit. Das Deutsche Text Archiv bietet u. a. die Möglichkeit Corpora mit dem Voyant Tool[27] zu analysieren (häufigste Terme und Phrasen, Häufigkeitenverteilung von Termen im Dokument, Volltextanzeige). Die Plattform der BYU Corpora bietet einige vorprozessierte Auswertungen für Einzelwörter (Kollokationen, Synonyme, Definitionen sowie Frequenzdistributionen).

Corpusplattformen, die auf der Abfragesprache CQP (Evert & Hardie 2011) und dem dafür von Andrew Hardie entwickelten webbasierten GUI CQPweb (Hardie 2012) basieren, vereinen eine effiziente und mächtige Abfragesprache mit verschiedenen Auswertungsmöglichkeiten (Häufigkeitsverteilungen, Kollokationsanalyse, Keywordanalyse). Die meisten Auswertungen sind nicht vorprozessiert, lediglich Frequenzlisten werden bereits bei der Corpusinstallation berechnet. Bei der Datenextraktion und Datenauswertung kann auf die gesamte Annotation des Corpus zugegriffen werden. Für die Extraktion von komplexen Datensätzen sind modulare Plattformen wie CQPweb besonders geeignet, da sie erlauben die Merkmale für die Extraktion sehr spezifisch zu definieren.

Ähnlich wie bei der Corpusexploration können generische Komponenten bei der Datensichtung und Datenevaluierung helfen, von den zugrundeliegenden Daten zu abstrahieren und so eine andere Sichtweise auf die extrahierten Daten ermöglichen, wodurch die Makrostruktur der Ergebnisse erst sichtbar wird. Generische Werkzeuge für die statistische Auswertung und Visualisierung wie R[28] aber auch Werkzeuge aus dem Data Mining (z. B. WeKa[29] oder Rapid Miner[30] für die Textklassifikation) bieten hier verschiedene Auswertungsmöglichkeiten, die je nach Untersuchungsgegenstand spezifisch angewandt werden können.

Durch den Einsatz von generischen Komponenten bei der Corpusanalyse kann diese effizient und schnell durchgeführt werden. Zudem gewährleistet der Einsatz von statistischen Verfahren und Methoden aus dem Data Mining ein hohes Maß an Objektivität. Es ist jedoch auch hier zu beachten, dass

---

27 http://voyant-tools.org/ (letzter Zugriff: 22. 4. 2018).

28 https://www.r-project.org/ (letzter Zugriff: 22. 4. 2018).

29 http://www.cs.waikato.ac.nz/~ml/weka/ (letzter Zugriff: 22. 4. 2018).

30 https://rapidminer.com/ (letzter Zugriff: 22. 4. 2018).

(i) automatische Werkzeuge keine perfekten Ergebnisse liefern, (ii) es nicht immer einfach ist, die Generalisierungen und Abstraktionen richtig zu interpretieren und (iii) die Werkzeuge per se keine spezifischen Analysen bieten. Dies bedeutet für spezifische Untersuchungen, dass bei der Interpretation der Blick auf die Mikroebene, also die Textgrundlage in Form von Textausschnitten, Konkordanzen oder Beispieldaten, nicht ausbleiben kann.

3 Zusammenspiel zwischen generischer Infrastruktur und spezifischer Forschung

In diesem Kapitel zeigen wir das Zusammenspiel zwischen generischer Infrastruktur und spezifischen Lösungen anhand eines konkreten Beispiels. Wir gehen zunächst kurz auf den linguistischen Hintergrund der Studie ein, beschreiben dann die Vorgehensweise bei der Corpuserstellung (Vorverarbeitung und linguistische Annotation) und gehen dann auf die Corpusanalyse (Merkmalsextraktion, Datensichtung und Datenevaluierung) ein.

Unser Interesse gilt der Entwicklung der englischen Wissenschaftssprache. Laut Halliday (1988) und Halliday & Martin (2005) kommt es hier aufgrund von Spezialisierung zu einer größeren Kodierungsdichte, d. h. kürzere, kompaktere sprachliche Ausdrücke werden häufiger benutzt, um dem Prinzip der Spracheffizienz zu entsprechen. Dabei zeigen sich Merkmale sprachlicher Verdichtung auf allen linguistischen Ebenen, z. B. Reduktion auf der syntaktischen Ebene (Artikelauslassung), Nominalisierungen auf der morphologischen Ebene und eine größere lexikalische Dichte auf der Wortebene (ausgeprägtere Verwendung von Inhaltswörtern).

Wir gehen nun davon aus, dass sich diese Art der Verdichtung am sprachlichen Signal als Informationsdichte messen lässt, also als Anzahl von Bits, die für die Kodierung einer gegebenen Äußerung notwendig ist (Shannon Information). Üblicherweise wird die Informationsdichte formal repräsentiert als die logarithmische Wahrscheinlichkeit einer sprachlichen Einheit gegeben einen Kontext (Crocker, Demberg & Teich 2015). Vereinfacht gesagt, je besser ein gegebener Kontext die sprachliche Einheit vorhersagen kann, desto kürzer ist die sprachliche Einheit (vgl. z. B. Variation in der Wortlänge, Mahowald et al. 2013) und desto weniger Bits werden für die Kodierung benötigt.

Für die Untersuchung brauchen wir einen Corpus des wissenschaftlichen Englisch, das eine gewisse zeitliche Ausdehnung hat, die Anfänge des wissenschaftlichen Schreibens in Englisch einschließt und eine möglichst breite Abdeckung im Bezug auf Disziplinen bietet. Vorhandene diachrone Corpora sind entweder auf eine Disziplin beschränkt oder decken nur eine bestimmte Zeit-

periode ab (z. B. das Corpus of Early Modern English Medical Texts, Taavitsainen & Pahta 2012) bzw. sind recht klein (z. B. das Coruña Corpus, Moskowich & Crespo 2007). Daher ist es hier nicht möglich auf eine vorhandene Ressource zurückzugreifen. Es musste ein Corpus neu erstellt werden. Betrachtet man die Rolle der *Royal Society of London* bei der Entwicklung der Wissenschaft ab Mitte des 17. Jahrhunderts (Atkinson 1998), so sind die *Philosophical Transactions* eine geeignete Datengrundlage. Die *Philosophical Transactions* wurden 1665 von Henry Oldenburg gegründet und sind die erste regelmäßig erscheinende Zeitschrift, die wissenschaftliche Artikel in englischer Sprache veröffentlichte. Sie enthielt ursprünglich sowohl wissenschaftliche Korrespondenz, Rezensionen und Zusammenfassungen von Büchern als auch wissenschaftliche Berichte von Beobachtungen und Experimenten. Als solches repräsentiert sie die Anfänge wissenschaftlichen Schreibens in englischer Sprache bis hin zu der Etablierung eines ersten wissenschaftlichen Standards.

3.1 Corpuserstellung

Ein weiterer Vorteil der *Philosophical Transactions* ist, dass sie bereits digitalisiert sind und von JSTOR zusammen mit Metadaten wie Autor, Texttyp und Jahr der Publikation etc. in wohlgeformtem XML bereitgestellt werden. Das Royal Society Corpus (RSC) umfasst alle Veröffentlichungen der *Philosophical Transactions* (Artikel, Buchrezensionen und Abstracts, sowie verschiedene Produktionsmodi, schriftsprachliche und gesprochensprachliche Texte) aus den ersten 200 Jahren der Zeitschrift von 1665–1869. In der aktuellen Version (2.0) hat das RSC ca. 35 Millionen Token. Bei den Digitalisaten handelt es sich um gescannte Texte aus unterschiedlichen Quellen (Tab. 5.1 zeigt einen Überblick über die Anzal der Texte pro Textkategorie und Zeitperiode), die noch weitgehend unerforscht bzw. unbekannt sind.

Tab. 5.1: Quellen des RSC.

		Buchrezen.	**Artikel**	**Misc.**	**Insg.**
Philosophical Transactions	1665–1678	124	641	154	919
Philosophical Transactions	1683–1775	154	3.903	338	4.395
Philosophical Transactions of the Royal Society of London	1776–1869	–	2.531	283	2.814
Abstracts of Papers Printed	1800–1842	–	1.316	15	1.331
Abstracts of Papers Communicated	1843–1861	–	429	5	434
Proceedings of RSL	1862–1869	–	1.476	38	1.528
Insgesamt		**278**	**10.296**	**833**	**11.421**

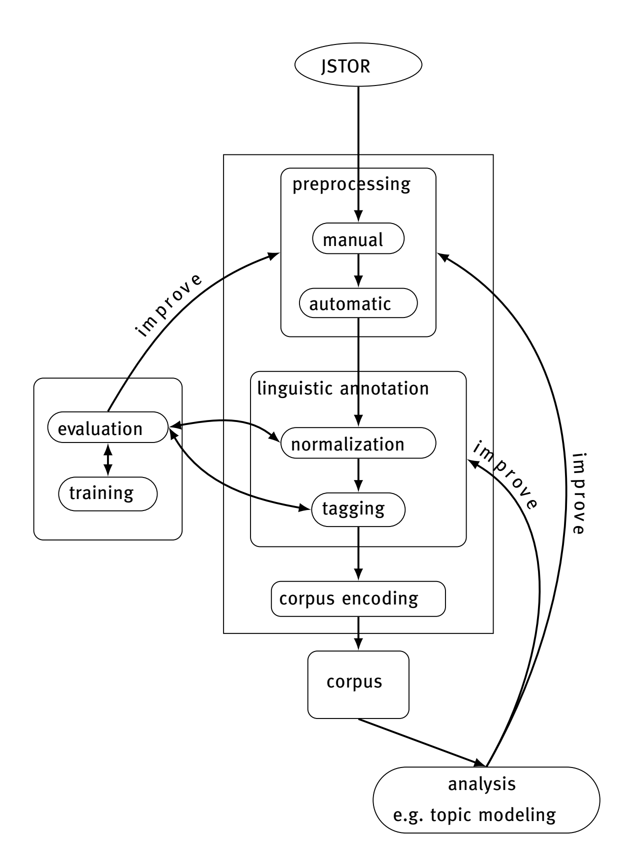


Abb. 5.4: Arbeitschritte beim Agile Corpus Building (vgl. Kermes et al. 2016a).

Wir haben uns daher für eine inkrementelle Vorgehensweise bei der Corpuserstellung entschieden, die sich an der Idee des *Agile Software Development* (Cockburn 2001) orientiert. Der gesamte Prozess der Corpuserstellung von den Ausgangsdaten bis zum abfragbaren Corpus ist weitgehend automatisiert und verwendet, wo immer möglich, generische Komponenten (etwa bei der linguistischen Annotation). Dedizierte automatische Komponenten ergänzen die generischen Werkzeuge, wenn eine generische Komponente nicht verfügbar ist. Manuelle Arbeitsschritte werden nur vorgenommen, wenn eine Automatisierung nicht möglich oder nicht sinnvoll ist und setzen direkt auf den Ausgangsdaten auf. Die Automatisierung hat den Vorteil, dass auf Probleme in der Datenqualität relativ schnell und effizient reagiert werden kann. Die entsprechenden Komponenten können angepasst oder ergänzt werden und es kann eine neue verbesserte Version des Corpus erstellt werden (cf. Kermes et al. 2016a, b). Abbildung 5.4 zeigt eine schematische Darstellung des Arbeitsablaufs.

3.2 Vorverarbeitung

Die Ausgangsdaten des RSC liegen zwar digitalisiert und in einem strukturierten XML-Format vor, eine Vorverarbeitung ist dennoch notwendig, um die Daten in ein standardisiertes Format zu überführen und zu bereinigen.

Bei der OCR-Fehlerkorrektur greifen wir auf die in Unterkapitel 2 *Methodik, Arbeitsabläufe und Angebote* beschriebene Ersetzungsliste von Underwood & Auvil (2012) zurück. Da das RSC jedoch recht spezifisch ist, zeigt sich, dass die Listen nicht einfach übernommen werden können, sondern an die speziellen Bedürfnisse angepasst werden müssen. Muster die für das RSC nicht relevant sind werden gelöscht, andere angepasst (so wird etwa ‚ſhe' zu ‚the' anstatt zu ‚she') und wieder andere Muster werden ergänzt. Für die Identifizierung der spezifischen Muster werden u. a. *word embeddings* verwendet, in der Annahme, dass falsch geschriebene Wörter ähnlich verwendet werden, wie das richtig geschriebene Wort. Bisher haben wir so ca. 360 spezifische Ersetzungsmuster ergänzt (cf. Knappen et al. 2017).

Bei der Lösung der Layoutprobleme muss ebenfalls auf spezifische Lösungen zurückgegriffen werden. Dedizierte Skripte kümmern sich etwa um Kopf- und Fußzeilen im Fließtext, in der Reihenfolge vertauschte oder fehlende Seiten, uneinheitliche Seitennummern, Seitenduplikate (erste und letzte Seite) und nicht eindeutig markierte Artikelgrenzen. Ist eine automatische Lösung nicht möglich, wird manuell oder (semi-)automatisch gesichtet. Bei einigen Quellen wurden Artikelgrenzen manuell annotiert. Texte und Seiten mit großen Tabellen wurden automatisch identifiziert und anschließend manuell gesichtet.

3.3 Linguistische Annotation

Wie oben bereits diskutiert, sind die meisten linguistischen Werkzeuge für gegenwartssprachliche, allgemeinsprachliche Corpora optimiert. Das RSC ist jedoch ein historisches Corpus aus wissenschaftlichen Texten. Es weicht also sowohl zeitlich als auch bezüglich des Registers ab. Trotzdem greifen wir bei der linguistischen Annotation auf generische Werkzeuge zurück, evaluieren die Ergebnisse und passen die Werkzeuge dann gegebenenfalls an.

Für die Normalisierung (hier Modernisierung) der historischen Originalwörter verwenden wir VARD (Baron & Rayson 2008), ein regelbasiertes statistisches Werkzeug, das orthographische Varianten bzw. historische Wortformen auf gegenwartssprachliche Wörter abbildet. VARD wurde für die Zeitperiode zwischen 1450–1700 entwickelt und überschneidet sich somit mit der Zeitperiode des RSC (1665–1869). Die Evaluierung von VARD zeigt eine Präzision von 61,8 %

und einen Recall von 31,4 %. Ein speziell trainiertes Modell verbessert die Präzision um mehr als 10 % auf 72,8 %. Der Recall verdoppelt sich auf fast 57,7 %.

Für Tokenisierung, Lemmatisierung und Wortartenannotation wird der TreeTagger (Schmid 1994; 1995) verwendet. Der TreeTagger ist ein Wortartentagger der auf gegenwartssprachlichem Zeitungstext trainiert wurde. Eine Evaluierung zeigt, dass der TreeTagger mit einer Präzision von 94 % (im Gegensatz zu 97 % auf gegenwartssprachlichen Texten) auch auf dem historischen Sprachmaterial zumindest akzeptable Ergebnisse erzielt. Eine detaillierte Analyse der Taggingfehler zeigt zwei Hauptfehlerquellen: NN-NP (Verwechslung von Nomen und Eigennamen) und WP-WDT (Verwechslung von Wh-Relativpronomen mit WhArtikeln). Beide Fehlerquellen sind für viele Analysen unproblematisch. Ignoriert man NN-NP Fehler, so erhöht sich die Präzision auf 95 %. Wir verwenden den TreeTagger daher im Augenblick fast unverändert. Lediglich das Lexikon des Tokenisierers wurde um ca. 170 Abkürzungen ergänzt. Dazu wurden zunächst Abkürzungskandidaten aus dem RSC extrahiert und die Häufigsten anschließend manuell gesichtet.

Für unsere Untersuchungen benötigen wir neben den klassischen linguistischen Annotation noch andere Informationen. So annotieren wir zusätzlich *Surprisal*, also den Informationsgehalt der Wörter in Anzahl Bits, berechnet als

$$S(unit) = -\log_2 p(unit|context)$$

d. h., der (negativen logarithmischen) Wahrscheinlichkeit einer gegebenen Einheit (z. B. eines Wortes) in einem Kontext (z. B. den vorangehenden Wörtern) (vgl. Genzel & Charniak 2002). *Surprisal* (der Informationsgehalt, Levy 2008) drückt die Intuition aus, dass je unwahrscheinlicher eine sprachliche Einheit in einem bestimmten Kontext ist, desto „überraschender“ (*more surprising*) oder informativer ist diese Einheit und desto mehr Bits werden benötigt, um sie zu kodieren (und umgekehrt). *Surprisal* erlaubt es, sprachliche Einheiten auf ihren Informationsgehalt hin zu untersuchen, den Kontext der Einheit bei der Untersuchung zu berücksichtigen und so über eine rein frequenzbasierte Untersuchung hinauszugehen. Für die Annotation von *Surprisal* gibt es bisher kein generisches Werkzeug, hier musste daher eine spezifische Lösung gefunden werden. Das dedizierte Skript annotiert *Surprisal* basierend auf verschiedenen Zeitperioden und erlaubt so den schnellen Zugriff auf diese Information. Obwohl hier als spezifische Lösung entwickelt, ist das Skript insofern auch generisch, als es auch auf andere Daten angewendet werden kann.

3.4 Beispielanalyse

Für die Beispielanalyse betrachten wir den Unterschied zwischen Funktionswörtern und Inhaltswörtern sowie deren Wortarten aus informationstheoretischer Sicht (Shannon 1949). Als Maß für den Informationsgehalt einer sprachlichen Einheit verwenden wir *Surprisal* und *Average Surprisal* (AvS), den durchschnittlichen Informationsgehalt. Dabei gehen wir von der Annahme aus, dass Funktionswörter generell besser vorhersagbar sind, ein niedrigeres AvS haben, während Inhaltswörter generell weniger vorhersagbar sind, ein höheres AvS haben. Außerdem nehmen wir an, dass das AvS von Funktionswörtern über die Zeit im Wesentlichen gleich bleibt, während das AvS von Inhaltswörtern weniger konstant ist.

Unsere Annahmen stützen sich auf bereits bekannte Unterschiede zwischen Funktionswörtern und Inhaltswörtern bezüglich Vorkommenshäufigkeit, Wortlänge, Anzahl (offene vs. geschlossene Wortklasse) und Informationsgehalt (cf. Biber et al. 1999). So zeigen Piantadosi, Tily & Gibson (2011), dass der durchschnittliche Informationsgehalt die Länge eines Wortes besser vorhersagt als dessen Häufigkeit. Laut Quirk et al. (1985: 72) ist die Anzahl der möglichen Wörter in typischen Kontexten von Inhaltswörtern größer als in typischen Kontexten von Funktionswörtern. Gleichzeitig zeigen Linzen & Jaeger (2015), dass die Anzahl an Ausdrucksmöglichkeiten die Vorhersagbarkeit der folgenden syntaktischen Konstruktion beeinflusst.

Beispielhaft schauen wir uns die Entwicklung im wissenschaftlichen Englisch an. Basierend auf der Annahme, dass es hier aufgrund der sprachlichen Verdichtung (cf. Halliday 1988; Halliday & Martin 2005) zu einer verstärkten Verwendung von Inhaltswörtern kommt, die oft durch lexikalische Dichte approximiert wird, nehmen wir an, dass wir diachrone Unterschiede beim AvS von Inhaltswörtern in wissenschaftlichen Texten beobachten können und dass diese Entwicklung sich von der Entwicklung in der Allgemeinsprache unterscheidet.

Für die Untersuchung verwenden wir das RSC und als Vergleichscorpus das *Corpus of Late Modern English Texts* (CLMET, Diller, De Smet & Tyrkkö 2011). Beide Corpora sind sowohl mit CQP[31] als auch mit dessen webbasiertes GUI CQPweb – also mit generischen Werkzeugen – abfragbar. CQPweb nutzen wir zur Corpusexploration und zum Aufbau und der Evaluierung der Abfrage. Dabei nützen wir den schnellen und flexiblen Zugriff auf Konkordanzen, Häufigkeitsverteilungen sowie Sortierungen und Gruppierungen der Ergebnisse, um die Adäquatheit unserer Abfragen zu überprüfen.

---

31 http://www.cwb.sourceforge.net (letzter Zugriff: 22. 4. 2018).

```
tabcmd: match word, match pos, match surpr50, match text_period
tabcmd_header: word, pos, surpr, time
ex: pos_avs:: [word = "\w+" & word !=".*[0-9_]+.*" &
               pos !="SYM|FW|UH|LS|CD|V[BH].*|N.*"] |
              [pos="N.*" & word =".{3,}" &
               word != ".*[0-9].*" & word = "\w+"]
ex: pos_avs_vbhaux:: [pos="V[BH].*"][pos!="VV.*"]{4}
ex: pos_avs_vbh:: [pos="V[BH].*"][]{0,3}[pos="VV.*"]
```

Abb. 5.5: Parameterdatei für die Merkmalsextraktion.

Für die Merkmalsextraktion nutzen wir dann das zugrundeliegende generische Abfragewerkzeug CQP. Es erlaubt einen Zugriff auf das Corpus durch externe Skripte. Durch die Automatisierung wird die Merkmalsextraktion reproduzierbar und auf andere Daten übertragbar. Die spezifischen Komponenten (Corpusabfrage und Merkmale) werden in Parameterdateien gespeichert, wobei eine Parameterdatei mehrere Abfragen enthalten kann. Mit einem dedizierten Extraktionsskript können dann für diese Parameter die entsprechenden Merkmale aus beliebigen CQP-Corpora extrahiert werden.

In unserem konkreten Fall wollen wir alle englischen Wörter aus dem RSC und dem CLMET extrahieren. Die Abfrage ist so formuliert, dass keine Fremdwörter, Symbole oder Zahlen extrahiert werden. Außerdem schließen wir Nomen aus, die aus weniger als drei Buchstaben bestehen. Die Merkmale, die wir für die Corpusinstanzen extrahieren sind das Wort selbst, die Wortart, der Surprisalwert und die Zeitperiode (50-Jahre-Zeitperioden). Die Verben *be* und *have* werden separat extrahiert, um zwischen Auxiliar und Vollverb zu unterscheiden. Abbildung 5.5 zeigt die Parameterdatei für die Extraktion.

Mit `tabcmd` und `tabcmd_header` werden die Corpusattribute und die Spaltennamen für die Merkmalsextraktion definiert. Die einzelnen Abfragen werden mit ex und einem Namen gekennzeichnet. Das Ergebnis der Extraktion ist eine TAB-getrennte Feature-Wert-Tabelle, mit einer Spalte für jedes Merkmal und einer Zeile für jede Corpusinstanz, die automatisch in einer Datei mit dem Namen der Abfrage gespeichert wird (s. a. Tab. 5.2).

Für eine bessere Abstraktion fügen wir der Tabelle zwei weitere Merkmale hinzu, indem wir die Wortartentags des verwendeten Tagsets (*Penn Treebank Tagset*, Marcus, Santorini & Marcinkiewicz 1993) in übergeordnete Wortarten sowie Funktionswörter (Artikel, Präpositionen, Pronomen, Modalverben, Konjunktionen und Auxiliarverben) und Inhaltswörter (Nomen, Adjektive, Verben, Adverben) gruppieren (s. a. Tab. 5.3).

Für die statistische Auswertung und Visualisierung verwenden wir R. R bietet einerseits den Zugriff auf bereits implementierte Auswertungen und

Tab. 5.2: Feature-Wert-Tabelle als Ergebnis der Extraktion.

word	pos	avs	time
An	DT	6.51	1650
Account	NP	1.12	1650
of	IN	0.06	1650
some	DT	2.66	1650
Books	NPS	0.90	1650

Tab. 5.3: Feature-Wert-Tabelle mit Abstraktion der Wortarten.

word	pos	apos	type	avs	time
An	DT	article	fw	6.51	1650
Account	NP	noun	cw	1.12	1650
of	IN	preposition	fw	0.06	1650
some	DT	article	fw	2.66	1650
Books	NPS	noun	cw	0.90	1650

andererseits die Möglichkeit eigene Funktionen für die spezifische Datenanalyse zu schreiben. Wir verwenden auch hier dedizierte Skripte. Dabei greifen wir einerseits auf bereits implementierte Auswertungen (z. B. den Mittelwert) und Visualisierungen (hier: Graphik zur Dichteverteilung) zurück und definieren andererseits spezifische Aspekte der Datenanalyse im Skript. Durch die Automatisierung mit dedizierten R-Skripten schaffen wir auch hier eine Reproduzierbarkeit der Ergebnisse. Durch Parameter (Datendatei, Corpus, Merkmale) in den R-Skripten sind die Analysen auch auf andere Daten übertragbar, z. B. auf Daten, die aus anderen Corpora extrahiert wurden oder auf anderen Extraktionen beruhen. Das Ergebnis ist dann z. B. die Graphik einer Dichteverteilung wie in Abbildung 5.6.

Die Abbildung zeigt die diachrone Entwicklung des AvS von Wortarten im RSC als Dichteverteilung. Wir sehen u. a., dass sich das AvS von einigen Wortarten im RSC über die Zeit tatsächlich verändert. So steigt das AvS von typischen modifizierenden Wortarten wie Adjektive, Adverbien, Modalverben sowie von Pronomen über die Zeit leicht an. Für Artikel, Präpositionen und Nomen sowie für Verben und Auxiliare sinkt das AvS.

Für einen Vergleich mit dem CLMET müssen wir nun dieselbe Merkmalsextraktion und Datenanalyse auf dem CLMET durchführen. Durch den Einsatz von generischen Komponenten und die Automatisierung und Modularisierung der spezifischen Skripte, ist eine Übertragbarkeit des Prozesses auf das CLMET

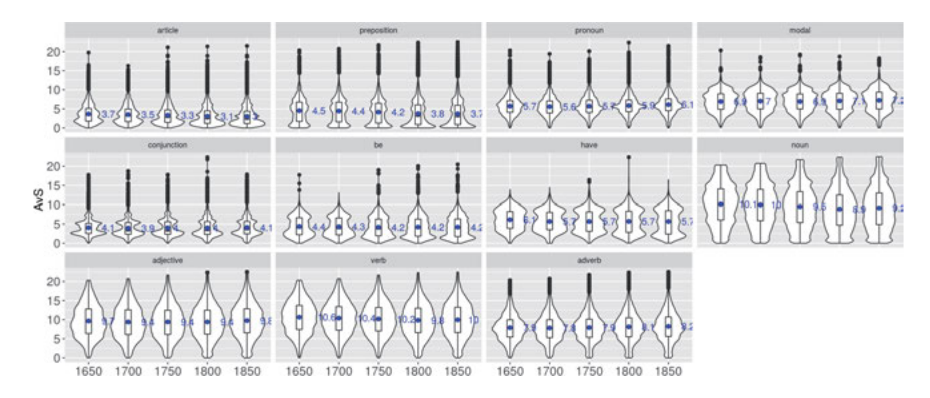


Abb. 5.6: Diachrone Entwicklung der AvS von Wortarten im RSC.

schnell und effizient möglich. Lediglich der Parameter des Corpus muss geändert werden (vgl. Kermes & Teich 2017) für eine ausführlichere Analyse des AvS von Wortarten im RSC).

4 Zusammenfassung und Schluss

Wir haben gezeigt welche Vorteile der Einsatz von generischen Infrastrukturkomponenten für spezifische Forschung haben kann: (i) Untersuchungen können auf größeren Datenmengen und effizienter durchgeführt werden und (ii) Ergebnisse können reproduziert und übertragbar gemacht werden. Dabei haben wir auch an einer konkreten Studie gezeigt, dass generische Infrastruktur auch spezifisch angepasst oder durch spezifische Lösungen ergänzt werden kann. Es zeigt sich, dass manche zunächst spezifischen Lösungen durchaus wiederverwendbar sind und so auch zu generischen Komponenten werden können bzw. diese ergänzen (z. B. erweiterte oder modifizierte Wortlisten für OCR-Korrektur, R-Skripte für komplexe Merkmalsextraktion).

Empirische Forschung an natürlichsprachlichen Daten kommt ohne den Einsatz von automatischen Verfahren, die über generische Werkzeuge (z. B. Tagger, Parser) zur Verfügung gestellt werden, nicht aus. Generische Werkzeuge unterstützen den Forschungsprozess und eröffnen neue Möglichkeiten, sie schließen aber spezifische Vorgehensweisen nicht aus und können auch manuelle Analyse und Interpretation nicht ersetzen. Dabei ist wichtig, dass man versteht, was die generischen Werkzeuge zu leisten im Stande sind und wo ihre Grenzen sind. Denn nur so ist gewährleistet, dass sie richtig eingesetzt werden und die Ergebnisse kritisch betrachtet und analysiert werden bzw. aus den Ergebnissen valide Schlussfolgerungen gezogen werden können.

Literatur

Atkinson, Dwight (1998): *Scientific discourse in sociohistorical context: The philosophical transactions of the Royal Society of London, 1675–1975*. Routledge.

Baron, Alistair & Paul Rayson (2008): VARD 2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*.

Biber, Douglas, Susan Conrad & Randi Reppen (1998): *Corpus linguistics: Investigating language structure and use*. New York: Cambridge University Press.

Biber, Douglas & Edward Finegan (2014): On the exploitation of computerized corpora in variation studies. In Karin Aijmer & Bength Altenberg (Hrsg.), *English Corpus Linguistics*, 204. Routledge.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan (1999): *Longman Grammar of Spoken and Written English*. Harlow, UK: Longman.

Brants, Thorsten & Oliver Plaehn (2000): Interactive Corpus Annotation. In *LREC*.

Buchholz, Sabine & Erwin Marsi (2006): CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, 149–164. Association for Computational Linguistics.

Burchardt, Aljoscha, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Pado & Manfred Pinkal (2006): SALTO–a versatile multi-level annotation tool. In *Proceedings of LREC 2006*, 517–520. Citeseer.

Cockburn, Alistair (2001): *Agile Software Development*. Boston, USA: Addison-Wesley Professional.

Crocker, Matthew W., Vera Demberg & Elke Teich (2015): Information Density and Linguistic Encoding (IDeaL). *KI – Künstliche Intelligenz* doi: 10.1007/s13218-015-0391-y.

Cunningham, Hamish, Diana Maynard & Kalina Bontcheva (2011): *Text processing with gate*. Gateway Press CA.

Cunningham, Hamish, Valentin Tablan, Angus Roberts & Kalina Bontcheva (2013): Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. *PLoS Computational Biology* 9(2), e1002854. doi: 10.1371/journal.pcbi.1002854.

Diller, Hans-Jürgen, Hendrik De Smet & Jukka Tyrkkö (2011): A European database of descriptors of English electronic texts. *The European English Messenger* 19, 21–35.

Düsendi, Bahadir (2014): *Erstellung annotierter Textcorpora mit WebLicht. Computerlinguistik als Sprachwissenschaft*. München: GRIN Verlag.

Eckart de Castilho, Richard, Chris Biemann, Iryna Gurevych & Seid Muhie Yimam (2014): WebAnno: A flexible, web-based annotation tool for CLARIN. In *Proceedings of the CLARIN Annual Conference (CAC)*.

Evert, Stefan & Andrew Hardie (2011): Twenty-First Century Corpus Workbench: Updating a Query Architecture for the New Millennium. In *Proceedings of the Corpus Linguistics 2011 Conference*, Brimingham, UK.

Genzel, Dmitriy & Eugene Charniak (2002): Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 199–206. Association for Computational Linguistics.

Halliday, M. A. K. (1988): On the Language of Physical Science. In Mohsen Ghadessy (Hrsg.), *Registers of Written English: Situational Factors and Linguistic Features*, 162–177. London: Pinter.

Halliday, M. A. K. & J. R. Martin (2005): *Writing Science: Literacy and Discursive Power*. Taylor & Francis.

Hardie, Andrew (2012): CQPweb – Combining Power, Flexibility and Usability in a Corpus Analysis Tool. *International Journal of Corpus Linguistics* 17(3), 380–409. doi: doi:10.1075/ijcl.17.3.04har.

Hinrichs, Erhard, Marie Hinrichs & Thomas Zastrow (2010): WebLicht: Web-based LRT services for German. In *Proceedings of the ACL 2010 System Demonstrations*, 25–29. Association for Computational Linguistics.

Kermes, Hannah (2008): Syntactic Preprocessing. In Anke Lüdeling & Merja Kytö (Hrsg.), *Corpus Linguistics. An International Handbook*, Band 1 Handbücher zur Sprach- und Kommunikationswissenschaft, 598–612. de Gruyter Mouton.

Kermes, Hannah, Stefania Degaetano, Ashraf Khamis, Jörg Knappen & Elke Teich (2016a): The Royal Society Corpus: From Uncharted Data to Corpus. In *Proceedings of the LREC 2016*, Portoroz, Slovenia.

Kermes, Hannah, Jörg Knappen, Ashraf Khamis, Stefania Degaetano-Ortlieb & Elke Teich (2016b): The Royal Society Corpus: Towards a high-quality corpus for studying diachronic variation in scientific writing. In *Proceedings of DH 2016*, Krakow, Poland.

Kermes, Hannah & Elke Teich (2017): Average surprisal of parts-of-speech. In *Proceedings of Corpus Linguistics 2017*, Birmingham.

Knappen, Jörg, Stefan Fischer, Hannah Kermes & Elke Teich (2017): The Making of the Royal Society Corpus. In *Proceedings of Nodalida 2017*, Göteborg.

Kübler, Sandra & Heike Zinsmeister (2015): *Corpus Linguistics and Linguistically Annotated Corpora*. Bloomsbury Academic annotated edition Ausg.

Langer, Hagen (2001): Syntax and Parsing. In Kai-Uwe Carstensen, Christian Ebert, Cornelia Endriss, Susanne Jekat, Ralf Klabunde & Hagen Langer (Hrsg.), *Computerlinguistik Und Sprachtechnologie. Eine Einführung*, 203–245. Heidelberg, Berlin: Spektrum Akademischer Verlag.

Leech, Geoffrey & Andrew Wilson (1996): EAGLES recommendations for the morphosyntactic annotation of corpora. *Version of March*.

Lemnitzer, Lothar & Heike Zinsmeister (2010): *Korpuslinguistik: Eine Einführung* NarrStudienbücher. Tübingen: Narr Verlag 2. Ausg. OCLC: 643072086.

Levy, Roger (2008): A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 234–243. Honolulu.

Linzen, Tal & T. Florian Jaeger (2015): Uncertainty and Expectation in Sentence Processing: Evidence From Subcategorization Distributions. *Cognitive Science*. doi: 10.1111/cogs.12274.

Mahowald, Kyle, Evelina Fedorenko, Steven T. Piantadosi & Edward Gibson (2013): Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition* 126(2), 313–318. doi: 10.1016/j.cognition.2012.09.010.

Manning, Christopher D. & Hinrich Schütze (1999): *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts and London, England: MIT Press.

Marcus, Mitchell P., Beatrice Santorini & Mary Ann Marcinkiewicz (1993): Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2), 313–330.

McEnery, Tony, Richard Xiao & Yukio Tono (2006): *Corpus-based language studies: An advanced resource book*. Taylor & Francis.

Moskowich, Isabel & Begoña Crespo (2007): Presenting the Coruña Corpus: A collection of samples for the historical study of English scientific writing. In Javier Pérez-Guerra &

Charles Jones (Hrsg.), *Of Varying Language and Opposing Creed': New Insights into Late Modern English*, 341–357. Bern: Peter Lang.

Müller, Christoph & Michael Strube (2001): MMAX: A tool for the annotation of multimodal corpora. In *In Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, Citeseer.

Piantadosi, S. T., H. Tily & E. Gibson (2011): Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences* 108(9), 3526–3529. doi: 10.1073/pnas.1012551108.

Quirk, Randolf, Sydney Greenbaum, Geoffrey Leech & Jan Svartvik (1985): *A comprehensive grammar of the English language*. London: Longman.

Schmid, Helmut (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, 44–49. Manchester, UK.

Schmid, Helmut (1995): Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*.

Schmid, Helmut (2008): Part-of-Speech Tagging. In Anke Lüdeling & Merja Kytö (Hrsg.), *Corpus Linguistics. An International Handbook*. de Gruyter Mouton.

Shannon, Claude E. (1949): *The mathematical theory of communication*. Urbana/Chicago: University of Illinois Press.

Taavitsainen, Irma & Päivi Pahta (Hrsg.) (2012): *Early Modern English Medical Texts. Corpus description and studies*. Amsterdam: John Benjamins.

Tognini-Bonelli, Elena (2001): *Corpus linguistics at work*, Band 6. John Benjamins Publishing.

Underwood, Ted & Loretta Auvil (2012): Basic OCR correction. http://usesofscale.com/gritty-details/basic-ocr-correction/ (letzter Zugriff: 12. 12. 2017).

Voutilainen, Atro (2003): Part-of-speech tagging. In Ruslan Mitkov (Hrsg.), *The Oxford Handbook of Computational Linguistics*, 219–232. Oxford University Press.

Yimam, Seid Muhie, Chris Biemann, Richard Eckart de Castilho & Iryna Gurevych (2014): Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 91–96. Baltimore, Maryland: Association for Computational Linguistics.

Yimam, Seid Muhie, Iryna Gurevych, Richard Eckart de Castilho & Chris Biemann (2013): WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. In *ACL (Conference System Demonstrations)*, 1–6.

Kerstin Eckart, Markus Gärtner, Jonas Kuhn und
Katrin Schweitzer

6 Nützlich und nutzbar für die linguistische Forschung: Sprachtechnologische Infrastruktur

Abstract: Dieser Beitrag soll veranschaulichen, wie die Gestaltung einer sprachtechnologischen Werkzeuginfrastruktur in der linguistischen Forschung von einer direkten wissenschaftlichen Relevanz sein kann, die signifikant darüber hinaus geht, dass – rein technisch motiviert – korpuslinguistische Forschung ermöglicht wird. Anhand von realen Szenarien aus der sprachwissenschaftlichen Forschung wird dargestellt, wie die Infrastruktur zur Einbindung sprachtechnologischer Komponenten als methodischer Rahmen dienen kann, innerhalb dessen sprachwissenschaftliche Erkenntnisse gewonnen werden können.

Keywords: Exploration, Infrastruktur, Korpora, Korpuslinguistik, Mündlichkeit, Visualisierung

---

Anmerkung: Die in diesem Beitrag angeführten Beispiele für Forschungsinfrastrukturkomponenten wurden im Rahmen mehrerer Projekte gefördert: im Rahmen des Sonderforschungsbereich 732 (Projekt INF) durch die Deutsche Forschungsgemeinschaft (DFG), im Rahmen von CLARIN-D durch das Bundesministerium für Bildung und Forschung (BMBF) und das Ministerium für Wissenschaft, Forschung und Kunst (MWK) Baden-Württemberg, im Rahmen von RePlay-DH durch das MWK Baden-Württemberg sowie im Rahmen von CRETA BMBF.

Kerstin Eckart, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart,
Pfaffenwaldring 5b, D-70569 Stuttgart, E-Mail: kerstin.eckart@ims.uni-stuttgart.de
Markus Gärtner, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart,
Pfaffenwaldring 5b, D-70569 Stuttgart, E-Mail: markus.gaertner@ims.uni-stuttgart.de
Jonas Kuhn, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart,
Pfaffenwaldring 5b, D-70569 Stuttgart, E-Mail: jonas.kuhn@ims.uni-stuttgart.de
Katrin Schweitzer, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart,
Pfaffenwaldring 5b, D-70569 Stuttgart, E-Mail: katrin.schweitzer@ims.uni-stuttgart.de

https://doi.org/10.1515/9783110538663-007

1 Einleitung

Dieser Artikel soll anhand von Beispiel-Szenarien zur Arbeit mit Mehrebenen-Annotation – insbesondere auf größeren mündlichen Korpora – verdeutlichen, wie sprachtechnologische Komponenten in den sprachwissenschaftlichen Forschungsprozess eingebunden werden können. Beispielsweise lassen sich prosodische Analysekomponenten, die auf akustischen Parametern des Sprachsignals basieren und durch automatische Annotation wesentlich größere Korpora erschließen können, mit Werkzeugen zur lexikalischen und syntaktischen Annotation zusammenführen. Durch Abgleich von Annotationsergebnissen, die auf unabhängigen Wegen ermittelt wurden, kann das Risiko von Annotationsfehlern reduziert werden. Eine in dieser Weise gewonnene ebenenüberspannende „Silberstandard“-Annotation macht beispielsweise explorative Studien zu prosodischen Realisierungsalternativen einer Gruppe von grammatischen Konstruktionen sehr leicht und erlaubt es, Korpusdaten in einer Größenordnung zu erschließen, die mit herkömmlichen Ansätzen nur unter extrem großem Aufwand angegangen werden könnten.

Der Infrastruktur, in der die Komponenten eingebettet sind, kommt in diesem Zusammenhang nicht allein die Funktion des technischen Rahmenwerks zu, sondern sie spielt eine eigene methodologische Rolle – unter anderem im linguistisch fundierten Prozess der Kombination von Teilergebnissen, in der Konfidenz-Abschätzung von approximativen Analysen, wie sie beim Einsatz von sprachtechnologischen Annotationskomponenten entstehen, sowie in der Qualitätsverbesserung von Annotationen durch wechselseitigen Abstimmung und Ausnutzung der jeweiligen Stärken von manuellen vs. automatisierten Annotationsschritten.

Hintergrund

Viele Zweige der sprachwissenschaftlichen Forschung legten von jeher großes Gewicht auf einen Zugang zu authentischen Sprachdaten – sei es um Varianz in der Lexik oder im grammatischen System empirisch zu beleuchten, sei es um Belege für diachrone Entwicklungen festzuhalten oder um mittels einer Konkordanz den Gebrauch bestimmter lexikalischer Ausdrücke unter die Lupe zu nehmen. Der Schritt hin zu elektronischen Korpora revolutionierte die Zugriffsmöglichkeiten auf die verfügbaren Sprachdaten, und seither ist die Korpuslinguistik eng eingebunden in die Entwicklung von geeigneten digitalen Ressourcen und Computerwerkzeugen, von Arbeitspraktiken, Workflows und validen Methoden für den Zugang zur sprachlichen Empirie in Form von Korpusevidenz. Einige wichtige Entwicklungen der Computerlinguistik und

Sprachtechnologie, wie der Einsatz und das Training von probabilistischen Analysemodellen (wie z. B. den Hidden-Markov-Modellen für das Part-of-Speech-Tagging, also der Auszeichnung von Wortarten), wurden durch den Bedarf und die Möglichkeiten der korpusbasierten Sprachforschung ausgelöst und befeuert. Am offensichtlichsten ist die gegenseitige Befruchtung bei der Automatisierung der Korpusannotation auf bestimmten Ebenen der linguistischen Beschreibung (Part-of-Speech-Tagging, Wortbedeutungs-Desambiguierung, syntaktische Analyse, Koreferenzanalyse) – andererseits liegen hier am ehesten auch die Angriffspunkte für eine kritische Haltung: Der unreflektierte Einsatz von automatischen Annotationswerkzeugen kann zu Ergebnissen führen, deren Validität nicht abgesichert ist. Richtig eingesetzt können sprachtechnologische Modelle jedoch sehr effektiv zur korpuslinguistischen Forschung beitragen (beispielsweise durch die Trennung von explorativen Schritten mit großem Skopus vs. streng operationalisierten Hypothesentests, für welche manuelle Absicherungsschritte hinzugefügt werden). Mit der laufenden Erschließung von weiteren linguistischen Ebenen für eine computergestützte Korpusannotation – beispielsweise an der Prosodie/GrammatikSchnittstelle – ist der Prozess der gegenseitigen Befruchtung weiterhin im Gange.

Vor dem Hintergrund der dargestellten Verwebung erscheint die Aussage, eine geeignete Infrastruktur für sprachtechnologische Werkzeuge (und für die Korpusressourcen selbst) sei von großer Bedeutung für die sprachwissenschaftliche Forschung, wenig überraschend. Wenn Filter- und Suchmethoden sowie diverse Modelle für eine automatische Analyse von Belang sind für eine korpuslinguistisch geprägte Sprachforschung, dann stellt sich – mittelbar – selbstverständlich auch die Frage nach den technologischen Rahmenbedingungen für die Vorhaltung, Ausführbarkeit und Ergebnisanalyse der Werkzeuge.

Der vorliegende Beitrag will darüber hinaus jedoch anhand einiger Szenarien deutlich machen, dass die Ausprägung der Infrastruktur – also verfügbare Ressourcen, Analyse- und Auswertungsmethoden etc. und ihre Kombinationsmöglichkeiten – bei der Erschließung von Korpusdaten für die sprachwissenschaftliche Forschung von zentralerer Bedeutung ist, als der Verweis auf einen generellen technischen Infrastrukturbedarf suggeriert: für Untersuchungen, die über gut erschlossene, homogene Subkorpora und etablierte Analysekategorien hinausgehen oder mehrere Ebenen der linguistischen Beschreibung berühren, kommt der Infrastruktur, in der die Analysekomponenten eingebettet sind, eine erweiterte Rolle zu: sie bildet den methodischen Rahmen, in dem sprachwissenschaftlich fundiert Teilanalysen zusammengeführt, die Adäquatheit möglicher Operationalisierungen eines Konstrukts überprüft oder mit indirekten Annäherungen an ein operationell schwer zu fassendes theoretisches Konzept experimentiert werden kann. Gewissermaßen erweitert die

Infrastruktur so die klassische Palette linguistischer Tests und fügt sich in eine differenzierte empirisch gestützte Argumentation ein.

Jenseits der „Standardkorpusdaten".
Wenn in der Korpusforschung der Untersuchungsgegenstand in einer oder mehreren Dimensionen von den gut erschlossenen Referenzkorpora abweicht bzw. wenn unterschiedliche Ebenen der linguistischen Beschreibung zusammengeführt werden, gelten bestimmte „Validitätsgarantien" aus dem Einsatz von Standardwerkzeugen auf Standarddaten nicht mehr: Das Part-of-Speech-Tagging eines vom orthographischen Standard abweichenden Untersuchungskorpus führt offensichtlich zu einer verringerten Vorhersagegenauigkeit gegenüber publizierten Evaluationsergebnissen auf Standard-Testdaten (die zunächst nicht genau beziffert werden kann). Folglich erzeugen Suchmuster für syntaktische Konfigurationen, die auf Part-of-Speech-Folgen aufsetzen, einen größeren Rauschanteil – wobei es sicherlich selbst unter der Hypothese einer perfekten Part-of-Speech-Analyse syntaktische Varianten gäbe, die durch das gewählte Suchmuster nicht erfasst werden (was bei der Auswertung der Filterergebnisse nach Möglichkeit berücksichtigt werden sollte). Selbst bei einer augenscheinlich „standardkonformen" Textgestalt führt der Einsatz von Werkzeugen außerhalb des Sprachregisters und der vorherrschenden inhaltlichen Textdomänen aus dem Entwicklungskorpus zu teils erheblichen Qualitätseinbußen.[1]

Nun wäre eine denkbare Reaktion auf diese Beobachtungen, automatische Annotationsschritte grundsätzlich nur dann einzusetzen, wenn das Untersuchungskorpus in allen relevanten Dimensionen eine strikte Erweiterung des Entwicklungskorpus darstellt – für welches eine sorgfältige Überprüfung der Werkzeugvalidität vorliegt. Eine Verlagerung auf andere Korpora wäre nur auf Basis einer Neuentwicklung oder umfassenden Anpassung der Modelle zulässig (einschließlich der manuellen Annotation einer ausreichend großen Stichprobe von Testdaten aus dem Zielkorpus). Als Konsequenz würde dieses Vorgehen allerdings die Vorzüge der Arbeit mit Computermodellen im Kern stark beschneiden: Um seltenere Varianten im Sprachgebrauch aufzufinden bzw. ihre Verwendungskontexte zu beleuchten, muss ein verhältnismäßig großes Korpus herangezogen werden, das in relevanten Dimensionen naturgemäß heterogen zusammengesetzt ist. Eine systematische Untersuchung eines entsprechend großen Korpus ist nur realistisch, wenn Computermodelle einen automatischen Zugriff erlauben. Das Risiko von falschen Modellvorhersagen lässt

1 Vgl. Sekine (1997).

sich folglich nicht im Vorfeld ausschließen. An dieser Stelle setzt jedoch der linguistisch motivierte Einsatz der Infrastruktur ein: Durch ein reflektiertes Vorgehen, eine geeignete Kombination von Modellkomponenten und gezielte Qualitätssicherungsmaßnahmen kann das Forschungsteam für eine gegebene Studie das Risiko, dass Vorhersagefehler zu Fehlschlüssen führen, stark eindämmen. Selbst ohne aufwändige Validierungsstudien – wie sie für die abschließende Absicherung der ganz zentralen Hypothesen selbstverständlich nötig bleiben – kann ein sorgfältig reflektierter explorativer Einsatz von Computermodellen in eine differenzierte Argumentation zur Eingrenzung der Forschungsfrage eingeflochten werden.

Dieser Beitrag will anhand von drei Szenarien zeigen, welchen Beitrag die Korpus- und Werkzeuginfrastruktur für ein solches Vorgehen leisten kann. In Abschnitt 2 stellen wir die sogenannte „Silberstandard-Methode“ der Korpusannotation vor, die unter anderem durch die Kombination von unterschiedlichen Annotationsverfahren einen linguistisch fundierten Korpuszugang erschließt. Beispielhaft ausgeführt werden Aspekte der Methode in einem Szenario, das ein mündliches Sprachkorpus mit einer Kombinationen aus automatischer syntaktischer Annotation, automatischer Annotation von prosodischen Parametern, sowie manueller Prosodie-Annotation auf einem Teilkorpus erschließt. In Abschnitt 3 diskutieren wir die Entwicklung von Explorationsverfahren für Korpora mit Mehrebenen-Annotation im Rahmen der interaktiven Plattform ICARUS, die u. a. zur Exploration und Visualisierung von mündlichen Korpora mit Annotation zur Dependenzsyntax, prosodischen Parametern und Koreferenz eingesetzt werden kann. Abschnitt 4 widmet sich Fragen des Workflows bei linguistisch komplexeren Korpusstudien, die mit sprachtechnologischen Komponenten arbeiten und weist auf die Bedeutung einer systematischen Buchführung über die zugehörigen Prozessmetadaten hin. Abschnitt 5 beschließt den Beitrag mit einer knappen zusammenfassenden Betrachtung.

2 Die Silberstandard-Methode

2.1 Quantität und Qualität von sprachtechnologisch verarbeiteten Daten

Ein wichtiger Schritt bei der sprachtechnologischen Verarbeitung von Daten ist die Annotation: das Explizitmachen von strukturellen Eigenschaften bestimmter Abschnitte in den Daten oder der Zuordnung der Abschnitte zu interpretato-

rischen Kategorien (im weitesten Sinne).[2] Dabei können die Abschnitte unterschiedlich groß sein und auf verschiedenen Ebenen annotiert werden. Zu einem mehrseitigen Dokument kann ein Thema annotiert werden, ein Abschnitt eines Dokuments kann als Kopfzeile gekennzeichnet werden, ein Abschnitt eines Satzes als Nominalphrase und ein Zeitpunkt in einer Audiodatei als Zielpunkt des Satzakzents bei der Realisierung der Satzmelodie.

Annotationen können manuell von Menschen oder automatisch von sprachtechnologischen Werkzeugen vorgenommen werden. Sie folgen sogenannten Annotationsrichtlinien, die auf theoretischen Überlegungen beruhen und mögliche Kategorien sowie Kriterien für deren Identifikation vorgeben (vgl. Lemnitzer & Zinsmeister 2015: 101 ff., zur Entwicklung von Annotationsschemata). Gängige linguistische Annotationsaufgaben sind in der Regel unter der Annahme einer weitgehend objektiven Entscheidungsmöglichkeit konzipiert. Das heißt, es wird davon ausgegangen, dass menschliche Annotatorinnen und Annotatoren nach einer Einweisungsphase aufgrund ihrer Sprachkompetenz und des im Korpus sichtbaren Kontexts in der Lage sind, intersubjektiv übereinstimmende Analyseentscheidungen zu treffen. Analysefragen, die individuell-subjektive Interpretationsentscheidungen beinhalten, sollten nach diesem Ansatz nicht in die Annotation eingehen. Während der Entwicklung von Annotationsrichtlinien werden entsprechend testweise manuelle Mehrfachannotationen desselben Materials durchgeführt. Wenn es zwischen den Annotierenden zu abweichenden Analyseentscheidungen kommt, muss geklärt werden, ob es sich schlicht um ein Versehen handelt oder ob die Richtlinien präzisiert werden können, damit im gegebenen Kontext intersubjektiv verlässliche Entscheidungen getroffen werden.[3]

Der Grundsatz der intersubjektiv entscheidbaren Annotationsziele erweist sich als sehr geeignete Arbeitshypothese – wenn sich auch nie restlos aus-

2 Neben sprachtechnologischen Werkzeugen, die für Eingabedaten in eine Ausgabestruktur überführen (unter Verwendung von Regelwissen und/oder probabilistischen Modellen), gibt es Werkzeuge, die inhärente Muster in den Daten mit „unüberwachten" Verfahren aufdecken. Beispiele dafür sind automatische Clustering-Verfahren für Wortformen oder andere Einheiten aufgrund des Verwendungskontexts und die *Latent Topic*-Modellierung (Verbreitet ist die Modellklasse der *Latent Dirichlet Allocation*, Blei, Ng & Jordan (2003).), mit der eine Textsammlung aufgrund der Kookkurrenz von (tendenziell semantisch verwandten) Wortformen nach automatisch induzierten Wortfeldern differenziert wird. Auch solche Ansätze können im sprachwissenschaftlichen Forschungsprozess gewinnbringend eingesetzt werden; hier konzentrieren wir uns jedoch auf Werkzeuge mit einer extern überprüfbaren Ausgabe – der Annotation.

3 Eine nicht unumstrittene Pointierung dieses Vorgehens liegt in der Strategie, die Annotationsaufgaben so zu wählen, dass ein 90 %iges Inter-Annotator-Agreement erreicht wird Hovy et al. (2006).

schließen lässt, dass gelegentlich die Annotation einer für sich genommen ambigen linguistischen Einheit (wie der Anbindung einer Präpositionalphrase bei der syntaktischen Annotation) auf impliziten Hinweisen aus dem Korpuskontext aufbaut – also eine subjektive Interpretation beinhaltet, in der auch kompetente Sprecher und Sprecherinnen zu unterschiedlichen Schlüssen kommen können. In der weit überwiegenden Zahl von Annotationsentscheidungen kann jedoch von einer intersubjektiv „korrekten" Option ausgegangen werden (in Bezug auf den Verwendungskontext) – dies ist eine wichtige Grundlage für die Entwicklung und Evaluation von automatischen Modellen und Werkzeugen.[4]

Der verbleibende Grad an Subjektivität in einer Annotationsaufgabe (kombiniert mit der Schwierigkeit, die Annotationsrichtlinien stets konsistent umzusetzen) lässt sich operationalisieren, indem dieselben Daten von mehreren gut trainierten Annotatoren und Annotatorinnen bearbeitet werden und die Übereinstimmung in den Annotationen, das „Inter-Annotator-Agreement", berechnet wird.

Unterschiedliche Annotationen können aufeinander aufbauen, zum Beispiel wenn zunächst eine Wortartenannotation erfolgt, die dann bei der Annotation von syntaktischen Strukturen berücksichtigt wird, oder verschiedene Sichtweisen auf die Daten wiedergeben, zum Beispiel syntaktische Phrasen und Intonationsphrasen.

Unabhängig davon, ob die Annotationen manuell oder automatisch erstellt werden, liegt das Ziel darin, die vorliegenden Daten möglichst genau entsprechend der Annotationsrichtlinien zu kategorisieren – also zu vermeiden, dass die Entscheidung von der „korrekten" Option abweicht (auf die sich eine Annotationsgruppe einigen würde, wenn sie eine ausführliche Diskussion führten). Unter Ausschluss des seltenen Falls, dass im gegebenen Kontext zwei alternative Annotationen gleichermaßen vertretbar sind, sprechen wir bei einer abweichenden Annotation von einem Annotations*fehler*.

Beim Vergleich des manuellen und des automatischen Annotationsansatzes wird deutlich, dass sich die jeweils typischen Arten von Fehlern unterscheiden. Schwierig und daher fehleranfällig für automatische Verfahren sind Entscheidungen, in denen gleichartige Eingabe-Konfigurationen unterschiedlich zu annotieren sind – in Abhängigkeit von linguistischem Wissen, das außerhalb der relevanten Beschreibungsebene liegt, bzw. sogar von Welt-

4 Damit wird nicht suggeriert, dass eine linguistische Einheit (als „Type") nicht systematisch ambig sein kann. Die Annahme ist jedoch, dass in jedem konkreten Verwendungskontext eine Desambiguierungsentscheidung getroffen werden kann. Das bewusste Spielen mit Mehrdeutigkeit, wie bei Ironie oder Wortwitz, wird hier ausgeklammert.

wissen. (Ein Beispiel sind globale syntaktische Ambiguitäten wie die Subjekt/Objektambiguität in „eine Entscheidung erwartet das Parlament noch heute“, die für menschliche Annotatoren zumeist offensichtlich auflösbar sind, während eine Operationalisierung von konsistenten, rein oberflächenabhängigen Kriterien unmöglich ist.) Automatische Werkzeuge können in diesen Fällen nur auf die vorgegebene Wissensbasis oder antrainiertes probabilistisches Wissen zurückgreifen, also eine Annotation „raten“. Bestimmte Zieloptionen, die sich aus dem Oberflächenkontext nicht erschließen lassen, werden diese Verfahren also systematisch falsch behandeln.[5] Gerade bei probabilistischen Ansätzen kann es bei nicht-trivialen Annotationsentscheidungen zu einem „Overfitting“ kommen, besonders wenn die Trainings- bzw. Entwicklungsdaten homogener waren als das Spektrum der Anwendungsdaten. Das Modell hat dann eine übermäßige Tendenz zu einer Annotationsentscheidung, die in den Trainingsdaten häufig auftrat. Beispiele gibt es auf allen Ebenen: So basieren die üblichen syntaktischen Trainingskorpora auf Zeitungsartikeln und enthalten beispielsweise nur sehr wenige Fragen. Für die Wortartenannotation führen Lemnitzer & Zinsmeister (2015: 58) (1), (2) und (3)[6] als Beispiele für unterschiedliche Lesarten von *einen* an. Während die Lesarten als indefiniter Artikel in (1) und Indefinitpronomen in (2) sehr wahrscheinlich vom Modell abgedeckt sind, ist die Verwendung als Verb wie in (3) unter Umständen unerwartet.

(1) Diese Perspektive ermögliche einen neuen Blick auf gesellschaftliche Verhältnisse.

(2) Gleichzeitig lautet der Appell an die Mieter, sich doch einen der Tiefgaragenplätze anzumieten.

(3) [Sie] wollen […] von Bremen aus die Republik wieder einen.

Bei der manuellen Annotation sind die eben genannten Fehlertypen praktisch nicht zu erwarten, dagegen sind hier Fehler häufig auf Ermüdungseffekte oder inkonsistenten Umgang mit feingliedrigen Unterscheidungen zurückzuführen. Dies schlägt sich so nieder, dass gleichartige linguistische Einheiten in vergleichbaren Kontexten inkonsistent annotiert sind. Fehler dieser Art lassen sich durch exhaustive Mehrfachannotation minimieren. Für anspruchsvolle manuelle Annotation ist ein Verfahren üblich, bei dem jede Abweichung in der Annotationsgruppe diskutiert wird oder von einer weiteren Person aufgelöst

5 Allerdings wird bei maschinellen Lernverfahren mit reichen Feature-Mengen das Verhalten nicht immer erkennbar systematisch sein.

6 (Gekürzte) Korpusbelege aus TüBa-D/Z; übernommen aus Lemnitzer & Zinsmeister (2015: 58).

wird. Die resultierenden Annotationen werden oft als „Goldstandard" bezeichnet, wobei sich „Gold" auf die zu erwartende hohe Qualität der Annotation bezieht, da diese nicht nur einer manuellen Annotation sondern auch einem Abgleich unterzogen wurden.

Vorteile der automatischen Annotation

Obgleich automatische Annotationen generell zu höheren Fehlerquoten führen, bringt die systematische Gleichartigkeit des Systemverhaltens Vorteile: Bei einer manuellen Nachkorrektur einer automatischen Annotation (im Ergebnis also einem semi-automatischen Verfahren) kann bei bestimmten Arten von Fehlern die Gesamtheit der möglichen Fehlerinstanzen im Korpus leicht ermittelt werden und dann systematisch nachannotiert werden (manche Fehler lassen sich auch praktisch automatisch korrigieren, sofern kontextsensitive Ambiguität ausgeschlossen werden kann).

Der größte Vorteil einer automatischen Annotation besteht jedoch darin, dass in kurzer Zeit weitaus größere Datenmengen verarbeitet werden können als bei der manuellen Annotation. Sofern die Korpusannotation vordringlich dazu dient, Belegstellen für eine anschließende (sprachwissenschaftliche) Feinuntersuchung aufzufinden, ist es – gerade bei der Erforschung niederfrequenter Phänomene – von übergeordnetem Interesse, über Annotationen eines möglichst großen Korpusausschnitts zu verfügen. Abstriche bei der Verlässlichkeit jeder Einzelannotation können eher in Kauf genommen werden. Für gängige Annotationsebenen, wie dem Part-of-Speech-Tagging oder der syntaktischen Dependenzstruktur, sind Datenmengen in der Größe von über 500 Millionen Sätzen, zum Beispiel aus Webkorpora (Schäfer 2015) verarbeitbar. Auf diese Weise können in der Tat auch für seltene strukturelle Phänomene Belegstellen aufgefunden werden (etwa Passive von reflexiven Verben wie: „Da wird sich gewunden und auf Zeit gespielt."[7]) – wenn sich auf Basis von Standardannotationen Suchfilter formulieren lassen, die die Belegkandidaten hinreichend einkreisen.[8]

Für Korpusstudien, die jenseits der explorativen Korpuserschließung Hypothesen zur quantitativen Verteilung von alternativen Varianten eines Phänomens oder konkurrierenden Konstruktionen empirisch überprüfen sollen, muss im Einzelfall validiert werden, ob ein eingesetztes automatisches Annotationsverfahren robust genug und hinreichend unabhängig von der Zielfrage ist,

---

7 Aufgefunden über Volltextsuche (Google Books) in: Claudius Rosenthal, Matthias Schreiber (2001): Nachdenkliches für Führungskräfte, Witten: SCM R.Brockhaus.

8 Vergleiche die Korpusstudie in Zarrieß, Schäfer & Schulte im Walde (2013).

so dass die studienrelevanten Frequenzzählungen durch die erwartbaren Annotationsfehler höchstwahrscheinlich nicht verfälscht werden. So dürfte es unbedenklich sein, sich auf die Vorhersagen eines automatischen Part-of-Speech-Taggers zu verlassen, wenn beispielsweise Subkorpora hinsichtlich eines stärker nominaler Schreibstils vs. eines stärker verbalen Schreibstils kontrastiert werden sollen (gegeben ein standardsprachliches Ausgangskorpus). Gelegentliche Fehler bei der Kategorisierung von Verben und Nomina sollten über die Subkorpora hinweg gleich wahrscheinlich auftreten. Hingegen könnte es problematisch sein, automatische Annotationen eines Dependenzparsers zur empirischen Vorhersage von abweichenden Valenzrahmen einer gegebenen Gruppe von niedrigfrequenten Verben heranzuziehen (womöglich auf Basis eines Sprachausschnitts, der sich von den Trainingsdaten des Parsers unterscheidet). In letzterem Fall besteht ein erhebliches Risiko, dass einige wenige Einzelfehler, wie sie bei jeder automatischen Annotation vorkommen, starke Effekte auf das intendierte Analyseergebnis haben.

In jedem Fall sollte das Analyseverhalten eines automatischen Annotationswerkzeugs unabhängig überprüft werden, sofern systematische Schlussfolgerungen aus der Annotation gezogen werden sollen. Bei Standard-Werkzeugen liegen in der Regel Performanzdaten für Goldstandard-Testdaten vor, die eine Größenordnung der erwartbaren Fehlerraten vorgeben. Allerdings ist bekannt, dass die Qualität von Annotationswerkzeugen sich zum Teil überraschend stark verschlechtert, wenn sich das Sprachregister, die Inhaltsdomäne und andere Faktoren in der Anwendung von den Entwicklungsdaten entfernen. Für eine verlässliche Einschätzung der Validität kann es daher notwendig werden, in die manuelle Annotation von Anwendungsdaten (entsprechend den aufgabespezifischen Annotationsrichtlinien) zu investieren und so einen eigenen Goldstandard für die relevanten Charakteristika der untersuchten Daten zu erzeugen.

Der Aspekt der Annotationsvalidierung in Bezug auf die Charakteristika des Anwendungskorpus ist in der Praxis häufig nicht im Bewusstsein der Sprachtechnologie-Anwender, wenn dank der leichten Verfügbarkeit von linguistischen Standard-Annotationswerkzeugen (beispielsweise im Rahmen der CLARIN-Infrastruktur[9] oder beim Einsatz der Stanford CoreNLPWerkzeuge[10]) Modelle ohne Absicht außerhalb der Entwicklungsdomäne und -subsprache eingesetzt werden. Problematisch kann dies insofern werden, als die möglichen Verfälschungen bis zuletzt übersehen werden können – sie werden in der Regel nur zu einer Verschiebung der Zwischenergebnisse führen, nicht zu radikal unerwarteten Verteilungen, die eher ins Auge stechen. Dennoch können

9 https://www.clarin.eu/ (letzter Zugriff: 16. 11. 2017).
10 Manning et al. (2014), https://stanfordnlp.github.io/CoreNLP/ (letzter Zugriff: 16. 11. 2017).

Schlussfolgerungen, die aus einer Analysekette gezogen werden, durch derartige Verfälschungen vollständig invalide werden.[11]

2.2 Zwischen Goldstandard und ungeprüfter Annotation

Da einerseits die Erstellung von Goldstandardannotationen ausgesprochen aufwändig ist, andererseits der ungeprüfte Einsatz von automatischen Annotationswerkzeugen Gefahren birgt, stellt sich bei der Arbeit mit großen, automatisch annotierten Korpora die Frage, ob und wie eine bessere Absicherung der Verlässlichkeit der Annotationen möglich ist. Zwischen den strengen methodischen Ansprüchen einer manuellen Goldstandardannotation (die jedoch aufgrund des von Natur aus schmalen Umfangs eingeschränkt nutzbar ist) und der unreflektierten *Ad-hoc*-Anwendung eines Sprachtechnologie-Werkzeugs auf das eigene Untersuchungskorpus (wie dies tausendfach geschieht) gibt es ein breites Spektrum von denkbaren Verfahren und Arbeitspraktiken, mit denen die effektive Annotationsqualität erheblich erhöht werden könnte. Eine entsprechend nachbereitete automatische Korpusannotation – ohne vollständige manuelle Überprüfung, nicht zuletzt wegen der Skalierbarkeit, sondern nach Möglichkeit mit zusätzlichen automatischen Verfahren – wird zuweilen als eine „Silberstandard-Annotation" bezeichnet (vgl. z. B. Rebholz-Schuhmann et al. 2010). Der Begriff lässt offen, welche Verfahren eingesetzt werden und wo im Spektrum das endgültige Ergebnis lokalisiert ist, eignet sich unseres Erachtens jedoch gut, um einen methodischen Anspruch für realistische reflektierte Korpusarbeit zu markieren. Die Autorinnen und Autoren, nicht zuletzt aus den Erfahrungen mit der projektübergreifenden Korpusarbeit im Rahmen des Sonderforschungsbereichs 732,[12] verstehen unter einem (dynamischen) „Silberstandardkorpus" daher auch keine statische Ressource, bei der am Ende

---

11 Ein Beispiel, das David Jurgens (Stanford University, Computer Science) anführt (Vortrag IMS, Universität Stuttgart, Oktober 2016) ist folgendes: Die Erkennung, in welcher Sprache eine Kurzmitteilung verfasst ist, gilt als gelöstes sprachtechnologisches Problem. Entsprechend werden beispielsweise nach Sprache gefilterte Twitter-Nachrichten für demographische Untersuchungen ausgewertet. Es zeigt sich allerdings, dass ein erheblicher Anteil von Kurznachrichten in afroamerikanischem Englisch nicht der Kategorie Englisch zugeordnet werden (was in den Standard-Testszenarien jedoch nicht ins Gewicht fällt). Nutzt man eine Analyse „aller" englischsprachigen Kurznachrichten beispielsweise für die Wahlforschung, kann es aber zu systematisch falschen Vorhersagen kommen.

12 SFB 732: Inkrementelle Spezifikation im Kontext; eine Kooperation zwischen dem Institut für Linguistik und dem Institut für Maschinelle Sprachverarbeitunge an der Universität Stuttgart; https://www.uni-stuttgart.de/linguistik/sfb732/ (letzter Zugriff: 16. 11. 2017).

einer Phase der automatischen Annotation einmalig bestimmte Konsistenz-Werkzeuge eingesetzt wurden und die Annotationen anschließend eingefroren wurden. Vielmehr sollten im Idealfall die Annotationen in einem Silberstandardkorpus sukzessive mehr und mehr abgesichert werden – indem beispielsweise eine Annotation in einem neuen Kontext genutzt wird und mögliche Inkonsistenzen zutage treten und korrigiert werden. Im Umkehrschluss heißt dies, dass zu einem gegebenen Zeitpunkt bestimmte Annotationsebenen im Silberstandardkorpus durchaus suboptimal sein dürfen – evtl. sogar weitgehend unmodifizierte automatische Analyseergebnisse, die jedoch als Untersuchungskontext für eine andere Ebene von Belang ist, deshalb mitgeführt und in der weiteren Entwicklung verfeinert wird. (Die Herkunft und der Verlässlichkeitsstatus sowie mögliche Workflows für eine systematische Verbesserung in Abhängigkeit von Weiterentwicklungen bei den verwendeten Komponenten sollten mit den Annotationen abgelegt werden. Diese verhältnismäßig komplexen Dokumentations- und Infrastrukturaufgaben diskutieren wir gezielt in Abschnitt 4.)

Im Sonderforschungsbereich 732 wird für die ebenenübergreifende linguistische Forschung ein verhältnismäßig großes Korpus von Radiointerviews aufgebaut und im Sinne der Silberstandard-Methode aufbereitet (Eckart & Gärtner 2016). Für die Interviews ist neben der Audio-Datei des Interviews jeweils ein orthographisches Transskript verfügbar, das bereits vom Radiosender (SWR2) angefertigt wird. Zwischen spontanen Dialogen und vollständig geplanten Redebeiträgen angesiedelt, eignen sich Radiointerviews sowohl als Testfeld für den Einsatz von automatischen Annotationswerkzeugen jenseits der computerlinguistischen Standard-Textsorte Zeitungsartikel als auch für die Anwendung von akustischen Modellen für die automatische Annotation von prosodischen Parametern, die bislang im Kontext des DIRNDL-Radionachrichtenkorpus entwickelt und getestet wurden.

Als Grundlage für eine Evaluation der automatischen Annotationsschritte auf dem Zielkorpus werden derzeit Goldstandardannotationen für einen kleinen Ausschnitt des Radiointerview-Korpus erzeugt, die Wortarten, Koreferenz und Diskursstruktur enthalten. Die Textdaten des Gesamtkorpus werden automatisch mit Wortarten und Dependenzstrukturen, die Audiodaten phonetisch und prosodisch annotiert. Die prosodische Annotation erfolgt mit Hilfe eines phonetischen Intonationsmodells, PaIntE, das im Folgenden näher beschrieben wird. Die Zusammenführung der Annotationen wird über die Positionsanker des „Linguistic Annotation Frameworks“ (ISO 24612:2012) ermöglicht. Als Explorationswerkzeug dient ICARUS (Gärtner et al. 2013: vgl. Abschnitt 3).

Größere Projektverbünde oder Kooperationen von Projekten mit verschiedenen theoretischen wie pragmatischen Ansätzen sind typische Anwendungs-

felder für die Silberstandardannotation. Denn gerade wenn verschiedene Sichten auf die gleichen Daten vorliegen und verschiedene Ebenen der linguistischen Betrachtung zueinander in Beziehung gesetzt werden, ergibt sich aus dem Konsistenzabgleich zwischen unterschiedlichen Annotationspfaden ein Informationsgewinn, der in die Exploration der Daten und die Konfidenzbewertung der Annotationen einbezogen werden kann (vgl. Abschnitt 2.3).

Automatische Annotation von Intonation mit PaIntE

Die Intonation einer Äußerung ist unter anderem durch „Intonationsereignisse" gekennzeichnet: lokale Minima oder Maxima im Grundfrequenzverlauf, die Prominenz hervorrufen. So kann zum Beispiel der Satz „Frau Merkel spricht" potentiell auf jedem der Wörter einen Satzakzent (typischerweise gekennzeichnet durch eine tonale Prominenz, Längung und größere Intensität) tragen. Die verschiedenen Realisierungen gehen mit verschiedenen Fokusstrukturen einher, so betont beispielsweise die Realisierung *FRAU Merkel spricht* (mit Satzakzent auf „Frau"), dass es sich um Frau, und nicht um Herrn Merkel handelt (kontrastiver Fokus). Solche Satzakzente können mithilfe des PaIntE-Modells (PaIntE: Parametrisierung von Intonations-Ereignissen; Möhler 1998; Möhler & Conkie 1998; Möhler 2001) modelliert werden.

PaIntE ist ein phonetischer, datenbasierter Ansatz zur Intonationsmodellierung, ursprünglich entwickelt zur Generierung des Grundfrequenzverlaufs (F_0Kontur) in der Sprachsynthese, von dem in jüngerer Zeit in verschiedenen korpusphonetischen Studien zur Satzintonation (Calhoun & A. Schweitzer 2012; Schauffler & K. Schweitzer 2015; K. Schweitzer et al. 2015) sowie in zwei Studien zu lexikalischen Tönen (Kelly & K. Schweitzer 2015; A. Schweitzer & Vu 2016) Gebrauch gemacht wurde. Darüber hinaus wurde das Modell erfolgreich in der automatischen Annotation von Prosodie eingesetzt (A. Schweitzer 2010). Das Modell passt an die im Signal gemessene (bzw. mit Standardverfahren geschätzte) und geglättete F_0-Kontur mithilfe einer parametrisierten mathematischen Funktion eine Kurve an, deren Verlauf von sechs freien Parametern bestimmt wird. Die Beschreibung der Kontur erfolgt über ein Fenster von drei Silben hinweg. Die Funktion modelliert vollständige Gipfel (also die Aufwärts- und die Abwärtsbewegung), aber auch Konturen, die lediglich steigen oder fallen, indem zwei sigmoide, also s-förmige, Kurven übereinander gelegt werden. Für Konturen ohne Gipfel kann auch nur ein Sigmoid, also zum Beispiel nur der Anstieg oder nur der Fall der Kontur, zur Modellierung verwendet werden. Das Modell entscheidet in mehreren Schritten, welches Modellierungsverfahren am besten zur tatsächlichen Kontur passt und welche Werte die sechs freien Parameter annehmen müssen, um die Original-Kontur am genauesten nachzubilden. Nach dem Modellierungsschritt ist also ein intonatorisches

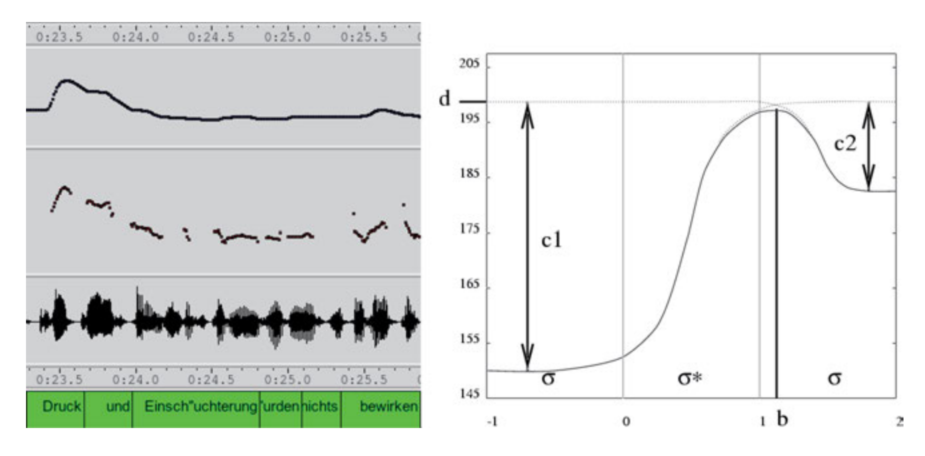


Abb. 6.1: Das PaIntE Modell. (**a**) zeigt einen Auszug aus dem DIRNDL Korpus, dargestellt in Wavesurfer (Sjölander & Beskow 2000). Die unterste Spur zeigt die Annotation auf Wortebene, darüber das Sprachsignal, gefolgt von der F_o-Kontur und in der obersten Spur wird die geglättete F_o-Kontur dargestellt. Das PaIntE-Modell operiert auf der geglätteten F_o-Kontur. (**b**) zeigt die PaIntE Modellierungsfunktion und ihre Parameter im drei-silbigen Analysefenster (Abbildung aus Möhler 2001). Die Silbe, für die die Parametrisierung ausgeführt wird, ist mit einem Stern gekennzeichnet. Auf der horizontalen Achse ist die Zeit (normalisiert über die Silbendauer) abgetragen. Die zwei sigmoiden Kurven bilden einen Gipfel in der dritten Silbe. Parameter b kodiert die zeitliche Verankerung des Gipfels. Auf der vertikalen Achse ist die Grundfrequenz in Hertz dargestellt. Die Größe des Anstieges vor, bzw. des Falls nach dem Gipfel entspricht den Parametern c1 und c2 – diese bezeichnen also Differenzen in Hertz. Parameter d reflektiert die absolute Höhe des Gipfels in Hertz. Parameter a1 und a2, die der Steigung der Kurve vor bzw. nach dem Gipfel entsprechen, sind nicht dargestellt.

Ereignis, zum Beispiel ein Satzakzent, durch sechs numerische Werte repräsentierbar.

Anders als andere Ansätze, wie zum Beispiel die automatische Beschreibung von Intonationskonturen mit Hilfe von Polynomen (vgl. Reichel 2014: und Referenzen darin), sind diese Parameter ohne weiteres linguistisch interpretierbar, da sie in naheliegendem Zusammenhang mit den üblichen perzeptiven Kategorien stehen – gleichzeitig sind sie theorieneutral, da komplexere theoretische Konzepte nicht gebündelt annotiert werden. So beschreiben zum Beispiel zwei Parameter (Parameter a1 und a2) die Steigung der Kurve vor bzw. nach dem Gipfel, ein weiterer Parameter (b) steht für die zeitliche Verankerung des Gipfels innerhalb des Analysefensters, das 3 Siben umfasst (die Silbendauer ist dabei normalisiert, so dass die aktuelle Silbe von 0 bis 1, und das gesamte Analysefenster von –1 bis 2 reicht), zwei Parameter (c1 und c2) beschreiben die Höhendifferenz im Anstieg vor dem Gipfel bzw. in der abfallenden Flanke

(dem „Fall“) nach dem Gipfel in Hertz und ein Parameter (d) kodiert die Höhe des Gipfels in absoluten Hertz-Werten (vgl. Abb. 6.1b).

Durch die Kodierung von Intonationsereignissen in direkt interpretierbaren numerischen Werten können diese sowohl mit Hilfe von etablierten numerischen Ähnlichkeitsmaßen verglichen als auch mit linguistischem Wissen klassifiziert werden (zum Beispiel kann das intonationsphonologische Wissen, dass bei einem fallenden Akzent der Anstieg vor dem Gipfel kleiner sein muss als der darauffolgende Fall, bei der Suche von fallenden Akzenten in die Bedingung $c1 < c2$ „übersetzt“ werden). Dadurch wird auch erreicht, dass Prosodie unabhängig von intonationsphonologischen Modellen auf phonetischer Ebene beschrieben werden kann.

Eine theorieunabhängige Beschreibung von Intonation bringt immense Vorteile im Bereich der Nachhaltigkeit und gemeinsamen Nutzbarkeit von Korpora gesprochener Sprache. So existiert zum Beispiel allein für das Deutsche eine Vielzahl etablierter intonationsphonologischer Modelle (Kohler 1991; Féry 1993; Mayer 1995; Grice, Baumann & Benzüller 2005; Peters 2005), die jeweils eigene Annotationsschemata benutzen, was die gemeinsame Nutzung von Korpora und anderen Ressourcen erschwert und nicht selten sogar unmöglich macht. Zwar gibt es in den letzten Jahren eine aktive Initiative, ein gemeinsames System für das Deutsche zu definieren (Kügler et al. 2015), dessen Annotationen problemlos in andere Systeme überführt werden können, jedoch ist diese nicht nur mit einem erheblichen Annotationsaufwand verbunden (der Zeitaufwand für das manuelle Annotieren von Intonation wurde mit 100- bis 200-mal der Echtzeit des Sprachsignals beziffert, vgl. Syrdal et al. 2001); sie erfordert zudem jeweils einen zusätzlichen Übersetzungsschritt. Zwar basiert ein automatisches Intonationsmodell wie PaIntE auf akustischen Eigenschaften, nicht auf menschlicher Perzeption und ist somit nicht äquivalent zu manuell annotierten Kategorien, jedoch kann ein automatischer Ansatz zur Datenexploration und zur Extraktion relevanter Subkorpora sowie zur Sichtung ungelabelten Materials und zur rein akustischen Analyse von Grundfrequenzverläufen verwendet werden. Durch den Einsatz eines Modells, das keine intonationsphonologischen Annahmen macht, werden Datensätze gesprochener Sprache theorie-übergreifend nutzbar – und dadurch noch nützlicher.

2.3 Maßnahmen zur besseren Nutzbarkeit großer annotierter Datenmengen

Eine Technik zur Absicherung bzw. Qualitätsverbesserung der Annotation im Rahmen der Silberstandard-Methode liegt in einem sogenannten extrinsischen Konsistenz-Abgleich, d. h., es sind zwei oder mehr Pfade hin zu einer bestimm-

ten Annotationsebene vorhanden, und die vorhergesagten Elemente auf dieser Ebene können miteinander verglichen werden (ohne Anschauung der intrinsischen Modelleigenschaften, die zur Vorhersage geführt haben – daher extrinsisch). Wir können unterscheiden zwischen einem horizontalen und einem vertikalen extrinsischer Konsistenzabgleich, wobei sich horizontal und vertikal auf die Relationen von Annotations- (bzw. Analyse-)Ebenen nach Eberle et al. (2012) beziehen.

Horizontaler Konsistenzabgleich

Hier werden mehrere alternative Annotationen auf derselben Analyseebene überprüft. Wurden diese mit unabhängigen Verfahren erzeugt, ist zu erwarten, dass die Fehler sich unterschiedlich verteilen. Ein Beispiel hierfür sind zum Beispiel mehrere Parser zur Erstellung syntaktischer Annotationen. George (2016) vergleicht die Ausgaben dreier statistisch trainierter Dependenzparser[13] und hält die Übereinstimmung als zusätzliche Annotationsebene fest, indem für die einzelnen Dependenzkanten der Grad des Konsens abgelegt wird – als Indikator für die Verlässlichkeit.[14]

Abbildung 6.2 zeigt die dependenzsyntaktische Analyse zweier Sätze durch den TurboParser[15] (Martins, Almeida & Smith 2013), trainiert auf einer Version des Nachrichtenkorpus TIGER[16] (Brants et al. 2004; Seeker & Kuhn 2012). Die Farbintensität entspricht der Übereinstimmung mit zwei weiteren Werkzeugen: Stimmen alle drei Parser bei der Wahl des syntaktischen Kopfes überein, ist die Kante dunkel – je heller die Kante, desto weniger Übereinstimmung. Die Sätze stammen aus Webdaten (Schäfer 2015), daher wird in Abbildung 6.2a der

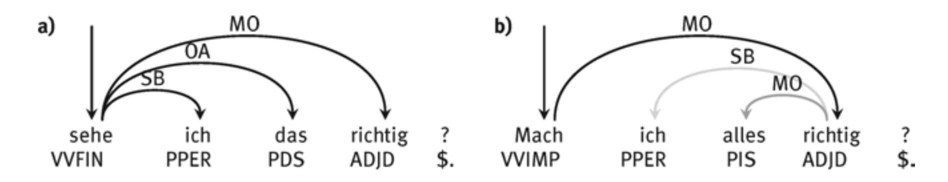


Abb. 6.2: Dependenzannotation mit Verlässlichkeitskodierung. Die Anbindung des Satzzeichens wird für den Vergleich im Beispiel außer Acht gelassen und ist daher nicht angegeben.

13 Graphbasierter Parser der Mate Tools (Bohnet 2010), IMSTrans (Björkelund & Nivre 2015) und TurboParser (Martins, Almeida & Smith 2013).

14 Manche Werkzeuge beinhalten auch die Abschätzung der Vorhersage-Konfidenz; diese könnte in vergleichbarer Weise in die Explorationsumgebung einbezogen werden.

15 http://www.cs.cmu.edu/~ark/TurboParser/ (letzter Zugriff: 16. 11. 2017).

16 Version 2.2.

Satzanfang kleingeschrieben und in Abbildung 6.2b die reduzierte Form *Mach* statt *Mache* verwendet. Für auf Nachrichtentexten trainierte Werkzeuge können solche Abweichungen von der erwarteten Schreibung ein Problem darstellen. Während die Abweichung in Abbildung 6.2a (satzinitiale Kleinschreibung) robust verarbeitet wird und alle drei Parser darin übereinstimmen, ist dies in Abbildung 6.2b nicht der Fall (wegen der abweichenden Schreibung wird der Form fälschlich das *Part-of-Speech*-Tag für Imperative zugewiesen, so dass der Dependenzparser ein Personalpronomen *ich* nicht als zugehöriges Subjekt integrieren kann). TurboParser erkennt das Verb zwar als Wurzel des Satzes und ordnet ihm den Modifikator *richtig* als Dependenten zu, weist jedoch dem Subjekt und Objekt nicht den korrekten Kopf zu. Nur bei den beiden korrekt erkannten Dependenzkanten stimmen die anderen beiden Parser mit TurboParser überein, wie aus den helleren Grautönen in den anderen Fällen deutlich wird – eine Einschränkung auf „Konsens-Kanten" bei der Korpussuche hätte in diesem Fall also die Verlässlichkeit erhöht. George (2016) evaluiert den Kombinationsansatz mithilfe des NoSta-D-Korpus (Dipper, Lüdeling & Reznicek 2013), das eine Goldstandard-Annotation von Dependenzanalysen auf verschiedenen Subkorpora enthält. Da sich die verwendeten Kategorien für die Dependenzrelationen zwischen NoSta-D und den Trainingsdaten der Parser unterscheiden, wurde die Evaluation auf einen Teil der Dependenzrelationen beschränkt. Als wichtige Annotationen wurden Hauptverb, Subjekt, Akkusativobjekt und Dativobjekt gewählt. Auf allen Subkorpora von NoSta-D (historisches Korpus, Chatkorpus, gesprochene Sprache, Lernerkorpus, literarische Prosa und Nachrichtentext) führt die Wahl der Konsens-Kanten zu einer Verbesserung der Genauigkeit (*Precision*), d. h. die als Annotation von Hauptverb, Subjekt, Akkusativobjekt und Dativobjekt identifizierten Kanten sind öfter richtig, als bei jedem der einzelnen Systeme; allerdings werden weniger der entsprechenden Kanten gefunden (*Recall*). Letzteres kann je nach Aufgabenstellung gerade auf sehr großen Datenmengen aber gegebenenfalls in Kauf genommen werden.

Generell deuten Fälle mit wenig Übereinstimmung oft auf schwierig zu annotierende Phänomene hin. Neben einem Ausschluss dieser Datenpunkte aus der Ergebnisliste zur Erhöhung der Verlässlichkeit kann auch gezielt nach diesen Fällen gesucht werden – beispielsweise um Bereiche zu identifizieren, für die eine systematische Nachkorrektur sinnvoll wäre (im gegebenen Beispiel könnte eine Modifikation des Part-of-Speech-Taggers angezeigt sein). Mit einer geeigneten Infrastruktur ist es denkbar, dass für Korpusrecherchen zu seltenen Phänomenen auf einem großen heterogenen Korpus sehr differenziert mit den „Eigenheiten" der unterschiedlichen Werkzeuge umgegangen werden kann.

Sogenannte Ensemble-Methoden nutzen die Kombination nicht nur zur Angabe von Verlässlichkeit, sondern erstellen beim Vergleich der einzelnen

Analysen eine kombinierte Annotationsebene, was im Ergebnis in der Regel zu einer höheren Annotationsqualität führt, da die unterschiedliche Fehlerverteilung der einzelnen Werkzeuge ausgenutzt werden kann (vgl. Surdeanu & Manning 2010: für Dependenzparsing).

Fasst man das Konzept der horizontalen Konsistenz etwas weiter, schließt es auch die Kombination von Werkzeugen ein, die zwar Annotation derselben linguistischen Beschreibungsebene erstellen, zum Beispiel Wortartenannotation oder Syntaxannotation, dies aber nicht auf der Grundlage der gleichen Annotationsrichtlinien bzw. theoretischen Frameworks tun. Haselbach et al. (2012) führen Informationen aus Dependenz- und Konstituentenstrukturen zusammen, um Argumentstrukturen von *nach*-Partikelverben verlässlicher aus Webdaten zu extrahieren als nur auf Basis einer der Analysen.

Vertikaler Konsistenzabgleich

Durch die Zusammenführung von Annotationen verschiedener Ebenen kann die Konsistenz von linguistischen Beschreibungen validiert und ggf. verbessert werden – und auf diese Weise kann für manuelle Annotationen, die relativ subtile Entscheidungen erfordern, eine Plausibilitätsprüfung implementiert werden. Bei der Kuration des DIRNDL-Korpus (Eckart, Riester & Schweitzer 2012) erfolgte dies durch einen Abgleich der phonetischen Intonationsbeschreibung durch das PaIntE-Modell mit der manuellen (perzeptiven) phonologischen Annotation von Intonation im GToBI(S)-Schema (Mayer 1995). Diese Art von extrinsischer Konsistenzüberprüfung auf der vertikalen Ebene wurde folgendermaßen vorgenommen: Die jeweilige Kategorie der manuellen Annotation wurde in eine phonetische Beschreibung unter Zuhilfenahme von PaIntE Parametern übersetzt. So ist zum Beispiel ein H*L Akzent kanonisch definiert als ein Satzakzent mit einem lokalen Gipfel in der akzentuierten Silbe, gefolgt von einem Fall in die postakzentuierte Silbe. Es wird angenommen, dass die F_0-Kontur zum lokalen Maximum hin ausgehend vom vorherigen Intonations-Ereigniss interpoliert wird, weswegen der Anstieg zum Gipfel hin weniger prominent ist als der Fall danach. Diese Beschreibung kann durch die PaIntE Parameter c1, c2 und b formalisiert werden: $c2 > c1$ fordert einen Fall der größer ist als der Anstieg, und $0 < b < 1$ fordert einen Gipfel in der akzentuierten Silbe. Anhand solcher Beschreibungen wurde die manuelle Annotation auf Plausibilität überprüft und gegebenenfalls manuell nachannotiert.

Eine weitere Möglichkeit, die Eigenschaften von automatischen Annotationen zur Qualitätsverbesserung auszunutzen, beruht auf der Gleichförmigkeit der Annotation über Korpusinstanzen hinweg, gegeben einheitliche Kontexteigenschaften. Wird also ein Annotationsfehler entdeckt, können sehr schnell systematisch gleichförmige Instanzen aufgerufen werden und möglicherweise

automatisch korrigiert werden (sofern keine Ambiguität vorhanden ist, die mit schwer zu fassenden Kontextfaktoren in Zusammenhang steht).[17]

Die bisherige Diskussion ging implizit von Annotationswerkzeugen aus, die für die jeweiligen Annotationskategorien die kontextuell (mutmaßlich) intendierte Ausspezifikation eventueller Ambiguitäten vorzunehmen – insbesondere also auch bei echter Ambiguität. Für synkretistische Formen, die auch im grammatischen Satzkontext formal nicht desambiguiert werden, wird also auf eine Präferenz aus möglichen Interpretation zurückgegriffen (dem gängigen Ansatz in der Sprachtechnologie folgend, für den eine Festlegung auf einzelne Lesarten naheliegend ist). Es gibt jedoch durchaus auch Annotationswerkzeuge, die bestehende grammatische Ambiguitäten als solche kennzeichnen, so der kaskadierte Finite-State Parser FSPar (Schiehlen 2003). (Viele andere bestehende Ansätze ließen sich mit einer entsprechenden Zielsetzung optimieren; im kombinierten morphologisch-syntaktischen Parsing-Ansatz von Seeker & Kuhn (2013) werden synkretistische Formen und Phrasen beispielsweise intern kodiert, die weitere Nutzung dieser Information wird jedoch nicht evaluiert, da in der Parser-Entwicklung sprachtechnologische Standard-Evaluationen etabliert sind.) Im Zuge germanistisch-sprachwissenschaftlicher Untersuchungen kann eine Einbeziehung dieser Information von großem Interesse sein – auch dies setzt eine ausdrucksstarke und handhabbare Infrastruktur zur Zusammenführung der Ergebnisse voraus.

Zusammenfassend wird mit der Silberstandard-Methodik also Information aus verschiedenen Quellen zusammengeführt, um automatische Annotationsansätze zu ergänzen, das Fehlerrisiko zu reduzieren und die Information besser nutzbar zu machen. Das führt nicht zwangsweise zu einer direkten Verbesserung der Annotationsqualität, aber zur verlässlicheren Exploration der Daten. Um die verschiedenen Datenebenen effektiv verwenden zu können, bedarf es einer Infrastruktur, welche die Forschenden bei der Exploration der vorhandenen Daten unterstützt ohne dabei die Forschung inhaltlich einzuschränken. Solche Infrastrukturen sind daher theorieneutral und können beliebige Kombinationen von Kategorien über verschiendene Ebenen abfragen. Sie müssen in der Lage sein, verschiedene Annotationsebenen sowie verschiedene Ausführungen einer Annotationsebene zu verarbeiten. Das umfasst auch und insbesondere die Schnittstelle zwischen Annotationen für gesprochene und geschriebene Sprache.

---

17 Ein sehr weitreichender Ansatz zur Qualitätsverbesserung von – vorwiegend manueller – Korpusannotation, die von der Gleichförmigkeit von Kontexten ausgeht ist der DECCA-Ansatz (Dickinson & Meurers 2003; Boyd, Dickinson & Meurers 2008), der auch innerhalb des ICARUS-Werkzeugs zur Verfügung gestellt wird (Thiele et al. 2014). Mit einem derartigen Verfahren wird die intrinsische Konsistenz eines Annotationsansatzes überprüft.

3 Exploration mit ICARUS

3.1 Eine interaktive Plattform zur Korpusanalyse

Nicht nur in der klassischen Korpuslinguistik und in der computerlinguistischen Forschung, sondern in sehr vielen anderen Bereichen der Linguistik und in textorientierten Disziplinen in den geistes- und sozialwissenschaften besteht ein wachsendes Interesse an der Arbeit mit annotierten Korpora. Der Umfang der verfügbaren Korpusressourcen und die Art und Menge enthaltener Annotationen wächst stetig an. In selbem Maße führt dies allerdings auch dazu, dass eine rein manuelle Untersuchung solcher Ressourcen ohne Unterstützung durch spezialisierte Werkzeuge immer weniger praktikabel ist.

Interaktive Anwendungen zur Visualisierung, Exploration und Suche stellen sicher, dass den Forschenden weiterhin ein einfacher Zugang zu den Inhalten großer Ressourcen gewährleistet werden kann. Nachfolgend wird ICARUS[18] als Beispiel einer solchen Anwendung vorgestellt und gezeigt, wie insbesondere Untersuchungen im Kontaktbereich von geschriebener und gesprochener Sprache von solchen interaktiven Zugangsmöglichkeiten profitieren können.

ICARUS entstand als Werkzeug zur Visualisierung und beispielbasierten Suche auf Dependenz-Baumbanken und wurde kontinuierlich erweitert, um sowohl Annotationen für Koreferenz (Gärtner et al. 2014), als auch Prosodie (Gärtner et al. 2015) zu unterstützen – immer unter Beibehaltung des Fokus auf einen einfachen und schnellen Zugang zu Korpus-Ressourcen. Im Gegensatz zu web-basierten Visualisierungs- und Such-Werkzeugen wie ANNIS (Krause & Zeldes 2014) ist ICARUS als eigenständige und lokal auszuführende Anwendung implementiert, hat aber trotzdem den Vorteil einer einfachen Einrichtung, die keine technischen Kenntnisse erfordert. Neben der Größe einzelner Korpora stellt auch die Vielzahl existierender Formate (CoNLL, TEI, TCF, Praat TextGrids, etc.), in denen diese verfügbar gemacht werden, Anwendungen vor immer wiederkehrende Herausforderungen. In der Basis-Version unterstützt ICARUS bereits diverse häufig verwendete Korpus-Formate und erlaubt auch, individuelle tabellarische Formate ähnlich den CoNLL-Formaten mittels eigener Schemata zu definieren. Zusätzlich ist die gesamte Anwendung modular als Sammlung von Plugins gestaltet, Erweiterungen zur Unterstützung anderer Formate oder Anpassungen bestehender Komponenten sind mit etwas Programmierkenntnis leicht umsetzbar. Da die Gesamtheit verfügbarer Annota-

---

18 *I*nteractive platform for *C*orpus *A*nalysis and *R*esearch tools, *U*niversity of *S*tuttgart (Gärtner et al. 2013).

tionen für viele Ressourcen eine übersichtliche Visualisierung schnell überladen würde, kann in ICARUS flexibel ausgewählt werden, welche Annotationsebenen wo und in welcher Form angezeigt werden sollen. Anwender und Anwenderinnen können somit die sichtbaren Informationen direkt an ihre individuellen Bedürfnisse anpassen.

Eine besondere Eigenschaft von ICARUS ist die Verknüpfung von Annotationen aus den sonst häufig getrennten Bereichen geschriebener und gesprochener Sprache, welche hier zusammen visualisiert, exploriert und abgefragt werden können. Ressourcen aus Phonetik und Phonologie sind typischerweise zeit-aligniert, Annotationen besitzen also Anker (sogenannte *time stamps*) auf spezifische zeitliche Abschnitte in einem Sprachsignal, zum Beispiel für die Markierung von Silben oder Phonemen. Dem gegenüber stehen Textkorpora mit Annotationen, die entweder auf Bereiche von Zeichen in einem Text (*character*-aligniert) oder bestehende linguistische Einheiten wie Wörter oder Sätze zeigen. Primärdaten in ICARUS werden grundsätzlich durch Text-Korpora repräsentiert, welche allerdings mit Audio-Dateien verknüpft werden können. Annotationen oder die Tokenisierung der Text-Daten können zusätzlich mit *Timestamp*-Informationen versehen sein, damit eine Zuordnung verschiedener Bereiche in geschriebener und gesprochener Sprache möglich wird.

In ICARUS führt dies dazu, dass Visualisierungen Koreferenz oder Dependenzsyntax gemeinsam mit Annotationen gesprochener Sprache anzeigen können, wie beispielsweise Approximationen der F_o-Kontur oder Betonungen auf einzelnen geschriebenen Wörtern oder Silben. Ebenfalls unterstützt wird das direkte Abspielen von Audiofragmenten für Textabschnitte auf verschiedenen Granularitätsebenen von Silben bis Sätzen. Für Forschende, die an Interaktionen von Phänomenen aus diesen beiden verschiedenen Modalitäten interessiert sind, entfällt somit die Notwendigkeit, mehrere Werkzeuge parallel verwenden zu müssen.

Die Kombination verschiedener Ebenen ist auch bei der Korpusabfrage in ICARUS möglich. Diese lässt sich direkt für quantitative Untersuchungen ausnutzen. Suchanfragen haben immer die Form einer (Baum-)Struktur, für die passende strukturelle Instanzen (Dependenz- oder Koreferenz-Strukturen) im jeweiligen Ziel-Korpus gesucht werden. Auf dieser Ebene unterstützt die Suche existenzielle Negation von (Kinds-)Knoten, Disjunktion, transitive Hüllen von Kanten-Erreichbarkeit und diverse Constraints für Knoten (Wurzel, Blatt etc.). Abhängig davon repräsentieren Knoten und Kanten in der Anfrage hierbei ebenfalls unterschiedliche Einheiten oder Konzepte und können zusätzlich zur grundlegenden strukturellen Definition der Suchanfrage mit beliebigen weiteren Constraints versehen werden, die dann zum Abgleich mit Inhalten entsprechender Kandidaten herangezogen werden. Für Suchen auf Syntaxbäumen sind

beispielsweise Wortart, morphologische Features oder Dependenz-Relation einzelne lokale Constraints. Für diese lokalen Constraints unterstützt ICARUS ein breites Spektrum an Operationen, unter anderem (Un-)Gleichheit, reguläre Ausdrücke und numerische Vergleiche. Speziell für quantitative Analysen und den Fall, ohne spezifische Kenntnisse der benutzten Annotationsschemata Korpusanfragen erstellen zu wollen, bietet ICARUS die Möglichkeit, für beliebige Constraints Instanzen und deren Häufigkeit ermitteln zu lassen und diese entweder in Frequenzlisten oder -Tabellen auszugeben. Mit einem gegebenen Korpus oder Annotationsschema wenig vertraute Anwenderinnen und Anwender können sich somit sukzessive detailliertere Informationen liefern lassen.

Die lokalen Constraints lassen sich dazu verwenden, linguistische Hypothesen und Theorien zu überprüfen, sie können aber auch für einen explorativeren Suchansatz benutzt werden, bei dem Forschende von einer beispielhaften Einzelinstanz ausgehend Muster in den Daten explorieren. Dies wird im folgenden Abschnitt dargestellt.

3.2 Beispielbasierte Suche mit ICARUS

Mit ICARUS können Korpora auch ohne detailliertes Wissen über die verwendeten Annotationsschemata zu durchsucht werden. Beispielbasierte Suche ermöglicht es Anwenderinnen und Andwendern, gezielt nach Realisierungen bestimmter Phänomene zu suchen, ohne deren Formalisierung in der jeweiligen Ressource zu kennen. Dies ist durch die Verwendung von Beispiel-Suchanfragen, also Suchanfragen, die eine konkrete Instanz der zu untersuchenden Struktur oder Realisierung beschreiben, möglich. Auf textueller Ebene kann durch Relaxierung der Suchanfrage vom Einzelbeispiel abstrahiert werden, auf der Ebene von gesprochener Sprache ist zusätzlich eine Generalisierung durch die Verwendung numerischer Ähnlichkeitsmaße möglich.

Im Folgenden werden beide Mechanismen anhand einer konkreten Korpusabfrage beschrieben.

Auf der Suche nach Bürgermeistern und Kanzlerinnen

Nehmen wir an, es soll die prosodische Realisierung von engen Appositionen (z. B. Lanwer, im Druck) untersucht werden. Hierfür wollen wir einen Überblick über die Realisierungen bekommen, ggf. ein bestimmtes oder mehrere bestimmte Betonungsmuster finden und ähnliche Beispiele im Korpus identifizieren. Als Beispiel-Korpus verwenden wir 1.624 Sätze aus DIRNDL (Eckart, Riester & Schweitzer 2012). Um herauszufinden, wie Appositionen im Korpus repräsentiert sind, kann zuerst ein Parse eines Beispielsatzes generiert werden. Beispielsweise liefert uns das Parsen des Satzes „Bürgermeister Meier lacht“ die gewünschte

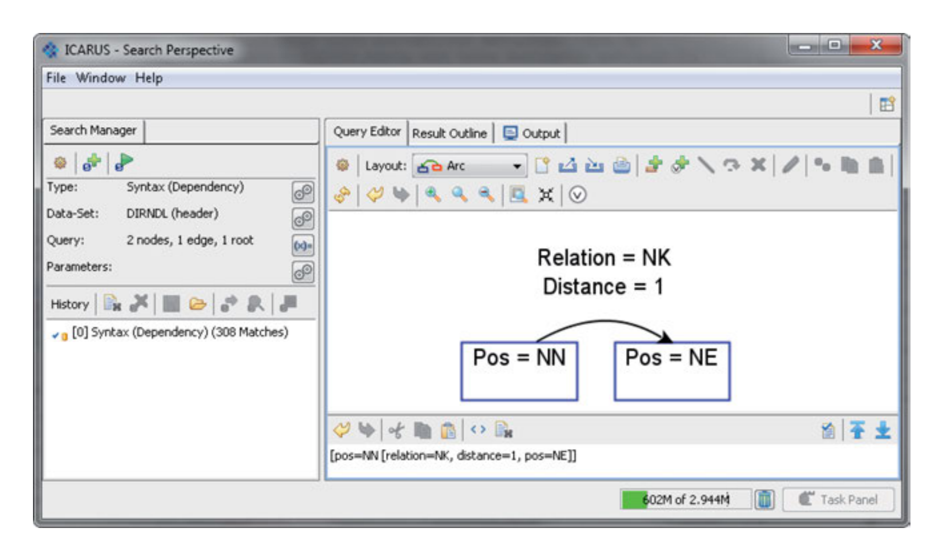


Abb. 6.3: Grafisches Interface für Korpusanfragen in ICARUS. Zu sehen ist eine Suchanfrage für enge Appositionen in grafischer (oberer Ausschnitt) und textueller Form (unteres Textfeld). In der Dependenzrelation Noun Kernel (NK) ist das normale Nomen (NN) der Kopf und der Eigenname (NE) der Dependent. Die Knoten sind direkt adjazent, haben also die Distanz 1.

Repräsentation. Diese Repräsentation kann nun als Referenz für die instanzbasierte Suche verwendet werden. Um ausgehend von dieser Referenz ähnliche Instanzen im Korpus zu finden, wird die Suchanfrage relaxiert – es wird von der konkreten Instanz abstrahiert. Dies kann zum Beispiel darüber erreicht werden, dass Information zu Lemmata und Wortformen verworfen und nur nach den entsprechenden Wortartenannotationen in der verwendeten syntaktischen Strutktur gesucht wird.

Abbildung 6.3 zeigt die relaxierte Suchanfrage für zwei adjazente Knoten (Distance = 1), die die Wortarten NN (normales Nomen) und NE (Eigenname) haben, wobei NN der Kopf und NE der Dependent einer Dependenzrelation NK (Noun Kernel) ist. Diese Konstruktion ist im Beispielkorpus 354-mal vorhanden, zum Beispiel *Außenminister Mottaki*, *Bundesaußenminister Steinmeier*, *Bundeskanzlerin Merkel*, aber auch in Form von *Seewetterdienst Hamburg*.

Abbildung 6.4 zeigt die Ergebnisliste mit einer Detail-Ansicht der prosodischen Realisierung.

In ICARUS ist es nun möglich, die Realisierung dieser Beispiele anzuhören, sowie die PaIntE Repräsentation in Form der Funktionskurve visuell zu inspizieren und schließlich zu extrahieren. Für die Exploration der gefundenen Treffer steht eine Vielzahl von vordefinierten Features zur Verfügung, mit der die Beispiele feiner klassifiziert werden können. Beispielsweise kann mithilfe des

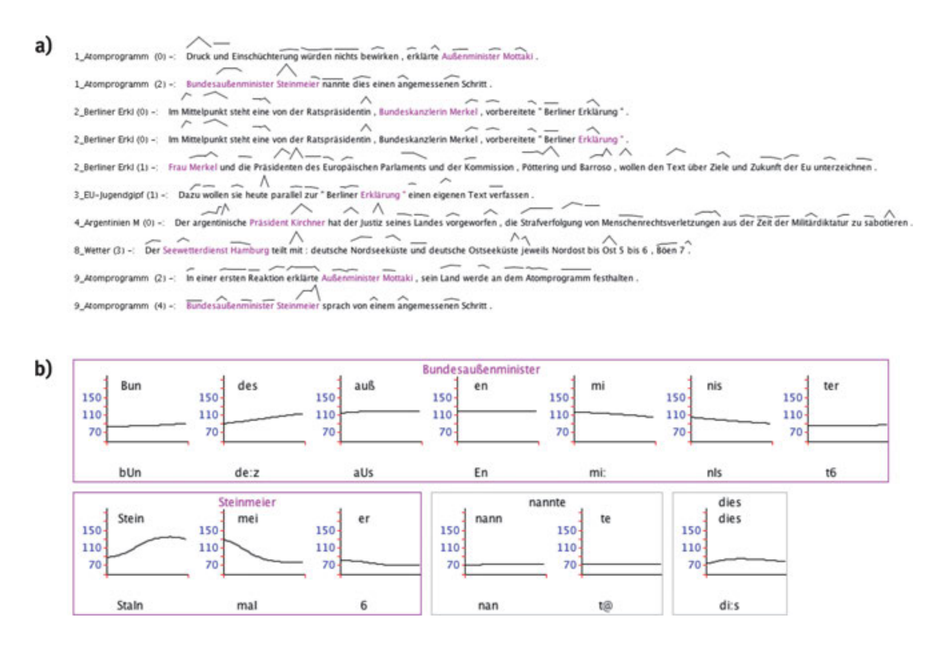


Abb. 6.4: Resultatsliste für enge Appositionen. (**a**) zeigt die Trefferliste. Die Treffer sind farbig markiert; die tonale Realisierung der Treffer wird in einer Voransicht dargestellt, die anhand der PaIntE-Parameter generiert wird. (**b**) zeigt die detaillierte Darstellung der PaIntE-Parameter eines Treffers.

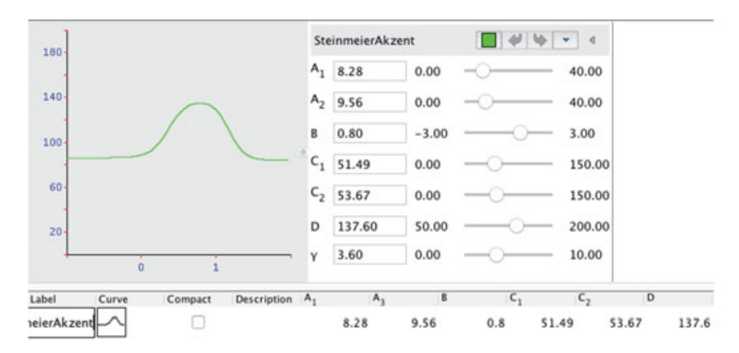


	false	true
false	253	68
true	21	12

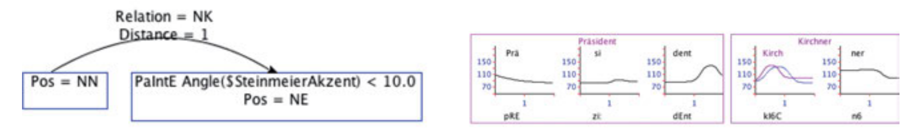
Abb. 6.5: Exploration der tonalen Realisierung von engen Appositionen. Das Feature „tonal prominence“ zeigt an, ob auf der betonten Silbe eines Wortes ein Gipfel mit einem Ausschlag von mindestens 50Hz realisiert wurde.

auf Wortebene operierenden Features „tonal prominence“ ein bestimmter Hertz-Wert definiert werden, der eine Mindesthöhe des lokalen Gipfels verlangt – somit können Grundfrequenzverläufe gefunden werden, die auf tonaler Ebene prominent markiert sind. Mittels eines Grouping Operators in der Suche kann eine Übersicht generiert werden, die zusammenfasst, welche Beispiele auf beiden Gliedern, auf einem oder auf keinem Glied der Apposition tonal prominent sind (vgl. Abb. 6.5).

Ebenfalls ist es möglich, auf Basis der PaIntE-Parameter definierte Gipfel-Formen zu finden – beispielsweise steigende oder fallenden Konturen. Darüber



a) Speicherfunktion für die Realisierung von *Steinmeier*



b) Suchanfrage nach Kosinus-Distanz

c) Ergebnisvivualisierung

Abb. 6.6: Ähnlichkeitssuche. Die tonale Realisierung von „Steinmeier" dient als Referenz. In der illustrierten Suchanfrage wurde zur Berechnung der Kosinus-Distanz der PaIntE-Parameter d, also die absolute Höhe des Gipfels, außer Acht gelassen, um einen Effekt von verschiedenen Stimmhöhen unterschiedlicher Sprecher auf die Berechnung der Distanzen zu vermeiden. Die Ergebnisse können gemeinsam mit der Referenzkontur angezeigt werden.

hinaus können verschiedene numerische Ähnlichkeitsmaße (Euklidische Distanz und Kosinus-Ähnlichkeit) angewandt werden, um von einer gefundenen Instanz wiederum zu abstrahieren, dieses Mal auf tonaler Ebene. Zum Beispiel ist der zweite Treffer (*Bundesaußenminister Steinmeier*) mit einem prägnanten Satzakzent markiert, dessen Realisierung wir nun als Prototyp für weitere Suchanfragen verwenden können. Die PaIntE-Parametersettings für diese Realisierung können gespeichert und direkt für eine feiner eingeschränkte Suche verwendet werden. Zur Exploration ähnlicher Realisierungen kann zum Beispiel als Ähnlichkeitsmaß die Euklidische Distanz oder die Kosinus-Distanz dieser Realisierung zu den anderen Übereinstimmungen im Korpus berechnet werden. Um einen Überblick über den Wertebereich der Distanzen zu unserem Zielakzent im Korpus zu bekommen, kann der Grouping-Operator und die Sortierungsfunktion verwendet werden. So können alle Suchtreffer nach Distanz zum Zielakzent sortiert und direkt visuell und auditorisch inspiziert werden. Im Anschluss kann ein Schwellenwert gesetzt werden, der nur Treffer bis zu einer gewissen Distanz anzeigen lässt. Abbildung 6.6 illustriert diese Vorgehensweise.

Es ist zu beachten, dass bisher in ICARUS noch keine Normalisierung der Parameterbereiche stattfindet, weswegen Parameter mit größeren Wertebereichen einen größeren Einfluss auf die Berechnung von akustisch-numerisch definierter Ähnlichkeit haben. Dies lässt sich momentan in gewisser Weise damit umgehen, dass das Ähnlichkeitsmaß nur für ein Subset der PaIntE-Parameter berechnet wird. Um alle Parameter mit einbeziehen zu können, kann eine Normalisierung der Parameterwerte, z. B. durch Z-Transformation, der Berechnung der Ähnlichkeit voraus gehen.

Um die für die jeweilige Forschungsfrage relevanten Korpus-Belege, die mithilfe der beispielbasierten Suche gefunden wurden, zu extrahieren, stellt ICARUS eine Exportfunktion bereit, deren tabellarisches Format zum Beispiel von gängigen statistischen Programmen wie R (R Core Team 2016) problemlos weiterverarbeitet werden kann.

4 Komplexe Abläufe reproduzierbar machen: Prozessmetadaten

Die Diskussion in den beiden letzten Abschnitten hat gezeigt, dass gerade ebenenübergreifende Korpusstudien und Studien, die heterogene Textsammlungen betreffen, davon profitieren können, dass nicht allein existierende Standardwerkzeuge eingesetzt werden, sondern dass Werkzeugen in geeigneter Weise kombiniert und angepasst werden.

Ein oft unterschätzter Aspekt bei der Kuration und Annotation von Ressourcen sowie bei Studien, die auf diesen Daten beruhen, ist die Reproduzierbarkeit der Abläufe, in denen mehrere Komponenten eine Rolle spielen. Ergebnisse können nur reproduziert und Studien auf mehreren Datensätzen nur dann verglichen werden, wenn die Schritte, die zu den Ergebnissen geführt haben, bekannt sind. Elming et al. (2013) zeigen anschaulich, wie ein Schritt in der Prozesskette, der nicht spezifisch für die Aufgabenstellung ist, Einfluss auf die Qualität der Ergebnisse haben kann: Es werden fünf verschiedene Aufgaben betrachtet, (i) die automatische Erkennung einer Negation und ihres Skopus, (ii) die Annotation semantischer Rollen, (iii) die maschinelle Übersetzung, (iv) die Satzkompression und (v) die Perspektivenklassifikation. Die Ausgangsdaten liegen mit einer Syntaxannotation in Form von Konstituentenbäumen vor und in der Prozesskette für jede Aufgabe werden die Konstituentenbäume in Dependenzstrukturen umgewandelt. Diese Umwandlung ist im Verhältnis zur jeweiligen Aufgabe nur ein kleiner Schritt, so dass ein großes Risiko besteht, dass dieser nicht in allen Einzelheiten dokumentiert wird (insbeson-

dere, wenn sich die eigentlich relevanten Ergebnisse auf weiterführende Aufgaben beziehen). Für eine Replizierbarkeit ist es jedoch außerordentlich wichtig, dass gerade auch die vermeintlich offensichtlichen Schritte exakt dokumentiert werden.

Elming et al. (2013) verwenden für jede Aufgabe vier verschiedene Konvertierungsprogramme, um die Umwandlung von Konstituenten in Dependenzstrukturen durchzuführen und zeigen damit, dass die Entscheidung, welche Konvertierung benutzt wurde, Einfluss auf die Ergebnisse der Studie hat. Um also die Ergebnisse richtig interpretieren zu können, ist es nicht nur notwendig zu wissen, dass die Konvertierung stattgefunden hat, sondern auch mit welchem Konverter.

Führt man diesen Gedanken weiter, bedeutet das aber auch, dass zur Reproduktion des Ablaufs auf neuen Datensätzen noch mehr Details relevant sein können. Zwischen Versionen von Programmen können Unterschiede bestehen, Programme können in verschiedenen Modi ausgeführt werden und greift das Programm auf weitere Wissensquellen, wie zum Beispiel Lexika, zu, spielt auch die verwendete Version des Lexikons eine Rolle. Bei statististischen Modellen, die auf unterschiedlichen Korpora trainiert werden können, sollte wenigstens über die Parameter-Datei – idealerweise über die exakte Version des annotierten Korpus und die Meta-Parameter des Trainingsprozesses Rechenschaft abgelegt werden.

Relevante Prozessketten bestehen jedoch nicht nur aus automatischen Verarbeitungsschritten sondern können auch manuelle Annotationen oder explizite Entscheidungen einschließen. Zum Beispiel spielt es eine Rolle bei der Erstellung eines Webkorpus, ob doppelte Vorkommen von Texten entfernt werden oder nicht. Gerade bei scheinbar offensichtlichen Zwischenschritten, die möglicherweise der üblichen Praxis folgen, besteht die Gefahr, dass sie nicht dokumentiert werden.

Die notwendigen Details zu dokumentieren bedeutet für die Forschenden zumeist einen nicht unerheblichen Mehraufwand neben der eigentlichen Forschungsarbeit, insbesondere da hierfür häufig noch keine dedizierten Werkzeuge zur Verfügung stehen. Infrastrukturen können das Sammeln der entsprechenden Prozessmetadaten und damit die strukturierte Dokumentation der Abläufe unterstützen, was nicht nur die Nachhaltigkeit der Ergebnisse und die Wiederverwendbarkeit der Ressourcen erhöht, sondern auch bereits im Forschungsprozess hilft, die Arbeitsschritte einzuordnen und ggf. auszutauschen.

Lösungsansätze aus anderen Disziplinen wie der Software-Entwicklung existieren bereits, sind aber nicht ohne weiteres zur Anwendung in den Sprachwissenschaften übertragbar. Beispielsweise kommen dort Versionie-

rungswerkzeuge wie Git,[19] SVN[20] oder Ähnliche zum Einsatz. Diese sind zwar in der Lage, schrittweise Änderungen der Inhalte von versionierten Dateien ausreichend detailliert zu verfolgen, allerdings überfordern sie zum einen Anwenderinnen und Anwender häufig durch ihre Komplexität oder zu viele Freiheiten, zum anderen bleibt die Problematik einer nachnutzbaren Prozessbeschreibung weiterhin ungelöst. Plattformen wie Kepler[21] und Galaxy,[22] unterstützen die direkte (Wieder)Ausführung von Workflows aus automatischen Schritten und haben meist einen Fokus auf naturwissenschaftlichen Anwendungen, z. B. in der Genetik. Typische Abläufe der Korpuslinguistik und ihrer Anwendung in den textverarbeitenden Geisteswissenschaften umfassen neben voll-automatischen Werkzeugketten auch semi-automatische Annotationen und manuelle Schritte, z. B. zur detaillierteren Bewertung und Auswahl von automatisch extrahierten Exemplaren.

Die Problematik wird im Kontext des koordinierten Forschungsdatenmanagements in Baden-Württemberg (bwFDM) mit dem Projekt RePlayDH[23] unmittelbar angegangen, mit dem Ziel, exemplarische Lösungsansätze zur Workflow-Dokumentation in den textverarbeitenden Geisteswissenschaften und der Computerlinguistik zu erarbeiten. Dabei soll die Dokumentation von Zwischenständen und einzelnen Arbeitsschritten bereits während des wissenschaftlichen Arbeitsprozesses erfolgen, mit der Option zur Archivierung, Veröffentlichung und Nachnutzung von Forschungsdaten.

Der gesamte Arbeitsablauf wird als Sammlung einzelner, getrennt aufgenommener Schritten modelliert, welche bestehende Ressourcen verändern/löschen oder neue hinzufügen. Im Hintergrund nutzt RePlay-DH hierbei die Funktionalität von Git, Übergänge zwischen Zuständen der Ressourcen mit FreitextBeschreibungen zu versehen, um formalisierte Prozessmetadaten zu hinterlegen. Inhalt dieser Prozessmetadaten sind Informationen über betroffene Ressourcen (Eingabe- und Ergebnisdateien), beteiligte Personen (Annotierende, Experiment-Beteiligte etc.), verwendete Werkzeuge (Parser, Tokenizer etc.) und deren Parameter/Konfigurationen, sowie eine freie textuelle Beschreibung des Arbeitsschrittes an sich. Bei der Angabe von Ressourcen kann unter anderem auf die breite Masse bereits verfügbarer Objekt-Metadaten aus eta-

---

19 https://git-scm.com/ (letzter Zugriff: 16. 11. 2017).

20 https://subversion.apache.org/ (letzter Zugriff: 16. 11. 2017).

21 https://kepler-project.org/ (letzter Zugriff: 16. 11. 2017).

22 https://galaxyproject.org/ (letzter Zugriff: 16. 11. 2017).

23 *Re*alisierung einer *Pla*ttform und begleitender Dienste zum Forschungsdatenmanagement für die Fachcommunit*y* Digital Humanities

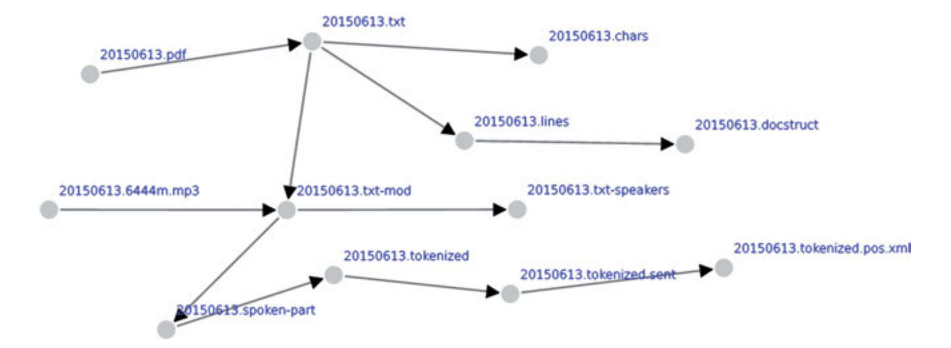


Abb. 6.7: Workflow-Visualisierung mithilfe der D3.js JavaScript Bibliothek. Es werden nur ausgewählte Arbeitsschritte visualisiert, die Dokumentation erfasst weitere Schritte und Details der verwendeten Werkzeuge.

blierten Repositorien (CLARIN Virtual Language Observatory,[24] TextGrid[25] etc.) zurückgegriffen und die entsprechenden Einträge in den Prozessmetadaten mit persistenten Links wie DOI[26] oder PID[27] versehen werden.

Ein besonderer Fokus des Projekts liegt darauf, die Erhebung der Prozessmetadaten so nicht-invasiv wie möglich zu gestalten, so dass sich diese ohne große Umgewöhnung oder Einarbeitung in bestehende Arbeitsabläufe integrieren lässt. Die im Rahmen von RePlay-DH entwickelte Software assistiert den Forschenden zu diesem Zweck bei der Erstellung oder Verlinkung von Prozessmetadaten und bietet umfangreiche Funktionen zur Visualisierung und Navigation des aktiven Workflows. Mit der Option, Metadaten zu einem Arbeitsprozess in einem zur Archivierung und Veröffentlichung geeigneten Format exportieren zu können, wird somit direkt die Nachnutzbarkeit von (Zwischen-) Ergebnissen ermöglicht, ohne dabei den sonst für Dokumentationen anfallenden und häufig unverhältnismäßig hohen Mehraufwand zu erzwingen.

Basierend auf ähnlichen Datenstrukturen werden Prozessmetadaten bei der Erstellung des in Abschnitt 2 beschriebenen Silberstandardkorpus festgehalten. Abbildung 6.7 zeigt die Visualisierung eines Workflowausschnitts bei

---

24 https://vlo.clarin.eu,\letzter\Zugriff:\16.\protect\kern+.1667em\relax11.\protect\kern+.1667em\relax2017.

25 https://textgrid.de/,\letzter\Zugriff:\16.\protect\kern+.1667em\relax11.\protect\kern+.1667em\relax2017.

26 Digitaler Objekt Identifikator https://www.doi.org/,\letzter\Zugriff:\16.\protect\kern+.1667em\relax11.\protect\kern+.1667em\relax2017.

27 Persistenter Identifikator https://www.clarin.eu/content/persistent-identifiers, \letzter\Zugriff:\16.\protect\kern+.1667em\relax11.\protect\kern+.1667em\relax2017.

dem ein Beispieldokument sowohl automatisch verarbeitet als auch manuell annotiert wird. Zunächst werden aus den, im PDF-Format vorliegenden, Interviews, Dateien im txt-Format erzeugt. Dann werden die Positionsanker zwischen den Zeichen verortet und pro Zeile zusammengefasst (Dateiendungen .chars und .lines). Danach werden Annotationen zur Dokumentstruktur erstellt (Kopfzeilen des Dokuments, welche Textbereiche sind die Beiträge der Interviewenden, welche die der Gäste; Dateiendung .docstruct). Für eine manuelle Wortartenannotation wird zunächst mithilfe der Audiodatei (Dateiendung .mp3) eine modifizierte Textversion erstellt (Dateiendung .txt-mod), bevor verschiedene Textteile (.spoken-part, .txt-speakers) exportiert und automatisch (Tokenisierung; .tokenized, .tokenized-sent) und manuell (Wortartenannotation; .tokenized.pos.xml) verarbeitet werden.

5 Abschließende Betrachtungen

Dieser Beitrag führte mehrere Szenarien aus der Arbeit mit digital aufbereiteten Sprach- und Textkorpora an, die auf technisch-infrastruktureller Ebene ein verhältnismäßig komplexes Zusammenspiel von Annotationskomponenten und Werkzeugen für die Abfrage und Exploration beinhalten: Verschiedene Ebenen der linguistischen Beschreibung müssen zueinander in Beziehung gesetzt werden – gerade wenn es um ebenenübergreifende Phänomene geht, etwa an der Prosodie/Grammatik-Schnittstelle. Bei automatischen Annotationsverfahren – gerade wenn sie jenseits der Standard-Zeitungskorpora eingesetzt werden – stellt sich die Frage der Zuverlässigkeit der Vorhersagen. Das Zusammenführen von unabhängig erzielten (Teil-)Analysen eröffnet andererseits Chancen für einen Abgleich, der für die Konsistenzüberprüfung ausgenutzt werden kann. Bei einem reflektierten Vorgehen kann die Kombination automatischer Annotationsverfahren den Zugang zu einem Vielfachen des Korpusumfangs erschließen, der mit herkömmlichen Techniken untersucht werden kann – gerade für mündliche Korpora, für die häufig nur kleine Ausschnitte mit manueller Annotation erschlossen sind.

Mit der Ausführung der Szenarien sollte verdeutlich werden, dass – sicherlich zunächst eher unerwartet – Fragen zur Gestaltung der Werkzeuginfrastruktur von direkter Relevanz für die germanistischen Sprachwissenschaft sein können – auch über die rein technische Motivation hinaus, dass eine geeignete Funktionalität für korpuslinguistische Forschungsansätze vorgehalten werden sollte.

Literatur

Björkelund, Anders & Joakim Nivre (2015): Non-deterministic oracles for unrestricted non-projective transition-based dependency parsing. In *Proceedings of the 14th International Conference on Parsing Technologies*, 76–86. Bilbao, Spain: Association for Computational Linguistics. http://www.aclweb.org/anthology/W15-2210 (letzter Zugriff: 12. 12. 2017).

Blei, David M., Andrew Y. Ng & Michael I. Jordan (2003): Latent Dirichlet allocation. *Journal of Machine Learning Research* 3, 993–1022.

Bohnet, Bernd (2010): Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 89–97. Beijing, China: Coling 2010 Organizing Committee. http://www.aclweb.org/anthology/C10-1011 (letzter Zugriff: 12. 12. 2017).

Boyd, Adriane, Markus Dickinson & W. Detmar Meurers (2008): On detecting errors in dependency treebanks. *Research on Language and Computation* 6(2), 113–137.

Brants, Sabine, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith & Hans Uszkoreit (2004): TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation* 2(4), 597–620.

Calhoun, Sasha & Schweitzer, A. (2012): Can intonation contours be lexicalised? Implications for discourse meanings. In Gorka Elordieta Alcibar & Pilar Prieto (Hrsg.), *Prosody and Meaning (Interface Explorations)*, 271–328. De Gruyter Mouton.

Dickinson, Markus & W. Detmar Meurers (2003): Detecting errors in part-of-speech annotation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, 107–114. Budapest, Hungary.

Dipper, Stefanie, Anke Lüdeling & Marc Reznicek (2013): NoSta-D: A corpus of German non-standard varieties. In Marcos Zampieri & Sascha Diwersy (Hrsg.), *Non-standard Data Sources in Corpus-based Research*, 69–76. Shaker.

Eberle, Kurt, Kerstin Eckart, Ulrich Heid & Boris Haselbach (2012): A tool/database interface for multi-level analyses. In *Proceedings of the Eighth Conference on International Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey: European Language Resources Association (ELRA).

Eckart, Kerstin & Markus Gärtner (2016): Creating silver standard annotations for a corpus of non-standard data. In Stefanie Dipper, Friedrich Neubarth & Heike Zinsmeister (Hrsg.), *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, Band 16 BLA: Bochumer Linguistische Arbeitsberichte, 90–96. Bochum, Germany. https://www.linguistics.rub.de/konvens16/pub/12_konvensproc.pdf (letzter Zugriff: 12. 12. 2017).

Eckart, Kerstin, Arndt Riester & Katrin Schweitzer (2012): A discourse information radio news database for linguistic analysis. In Christian Chiarcos, Sebastian Nordhoff & Sebastian Hellmann (Hrsg.), *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*, 65–75. Heidelberg: Springer.

Elming, Jakob, Anders Johannsen, Sigrid Klerke, Emanuele Lapponi, Hector Martinez Alonso & Anders Søgaard (2013): Down-stream effects of tree-to-dependency conversions. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 617–626.

Atlanta, Georgia: Association for Computational Linguistics. http://www.aclweb.org/anthology/N13-1070 (letzter Zugriff: 12. 12. 2017).
Féry, Caroline (1993): *German intonational patterns*. Tübingen: Niemeyer.
Gärtner, Markus, Anders Björkelund, Gregor Thiele, Wolfgang Seeker & Jonas Kuhn (2014): Visualization, search, and error analysis for coreference annotations. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 7–12. Baltimore, Maryland: Association for Computational Linguistics. http://www.aclweb.org/anthology/P14-5002 (letzter Zugriff: 12. 12. 2017).
Gärtner, Markus, Katrin Schweitzer, Kerstin Eckart & Jonas Kuhn (2015): Multi-modal visualization and search for text and prosody annotations. In *Proceedings of ACLIJCNLP 2015 System Demonstrations*, 25–30. Beijing, China: Association for Computational Linguistics and The Asian Federation of Natural Language Processing. http://www.aclweb.org/anthology/P15-4005 (letzter Zugriff: 12. 12. 2017).
Gärtner, Markus, Gregor Thiele, Wolfgang Seeker, Anders Björkelund & Jonas Kuhn (2013): ICARUS – an extensible graphical search tool for dependency treebanks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60. Sofia, Bulgaria: Association for Computational Linguistics. http://www.aclweb.org/anthology/P13-4010 (letzter Zugriff: 12. 12. 2017).
George, Tanja (2016): *Confidence estimation for automatic parsing of large web data sets*. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart Masterarbeit.
Grice, Martine, Stefan Baumann & Ralf Benzüller (2005): German intonation in autosegmental-metrical phonology. In Sun-Ah Jun (Hrsg.), *Prosodic Typology. The Phonology of Intonation and Phrasing*, 55–83. Oxford: Oxford University Press.
Haselbach, Boris, Kerstin Eckart, Wolfgang Seeker, Kurt Eberle & Ulrich Heid (2012): Approximating theoretical linguistics classification in real data: The case of German "nach" particle verbs. In *Proceedings of COLING 2012*, 1113–1128. Mumbai: The COLING 2012 Organizing Committee. http://www.aclweb.org/anthology/C12-1068 (letzter Zugriff: 12. 12. 2017).
Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw & Ralph Weischedel (2006): Ontonotes: The 90 % solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, 57–60.
Kelly, Niamh & Schweitzer, K. (2015): Examining lexical tonal contrast in Norwegian using intonation modelling. In *Proceedings of the 18th International Congress of Phonetic Sciences*, 1–5. Glasgow, UK. Paper number 0079.
Kohler, Klaus J. (1991): A model of German intonation. *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung (Univ. Kiel), AIPUK* 25, 295–360.
Krause, Thomas & Amir Zeldes (2014): Annis3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities* doi: 10.1093/llc/fqu057. http://dsh.oxfordjournals.org/content/early/2014/12/02/llc.fqu057 (letzter Zugriff: 12. 12. 2017).
Kügler, Frank, Bernadett Smolibocki, Denis Arnold, Stefan Baumann, Bettina Braun, Martine Grice, Stefanie Jannedy, Jan Michalsky, Oliver Niebuhr, Jörg Peters, Simon Ritter, Christine T. Röhr, Schweitzer, A., Schweitzer, K. & Petra Wagner (2015): DIMA – Annotation guidelines for German intonation. In *Proceedings of the 18th International Congress of Phonetic Sciences*, 1–5. Glasgow, UK. Paper number 0317.
Lanwer, Jens Philipp (2017): Apposition: A multimodal construction? The multimodality of linguistic constructions in the light of usage-based theory. *Linguistics Vanguard. A Multimodal Journal for the Language Sciences*. 3(1), 1–12. doi: 10.1515/lingvan-2016-0071.

Lemnitzer, Lothar & Heike Zinsmeister (2015): *Korpuslinguistik – Eine Einführung* narr Studienbücher. Tübingen, Germany: Narr Francke Attempto Verlag 3. Ausg.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard & David McClosky (2014): The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, 55–60. http://www.aclweb.org/anthology/P/P14/P14-5010 (letzter Zugriff: 12. 12. 2017).

Martins, Andre, Miguel Almeida & Noah A. Smith (2013): Turning on the turbo: Fast third-order non-projective turbo parsers. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (volume 2: short papers)*, 617–622. Sofia, Bulgaria: Association for Computational Linguistics. http://www.aclweb.org/anthology/P13-2109 (letzter Zugriff: 12. 12. 2017).

Mayer, Jörg (1995): Transcription of German intonation. The Stuttgart System. Unpubliziertes Manuskript. http://www.ims.uni-stuttgart.de/phonetik/joerg/labman/STGTsystem.html (letzter Zugriff: 12. 12.2017).

Möhler, Gregor (1998): Describing intonation with a parametric model. In *Proceedings of ICSLP 98*, Band 7, 2851–2854.

Möhler, Gregor (2001): Improvements of the PaIntE model for F_0 parametrization. Tech. rep. Institute of Natural Language Processing, University of Stuttgart. Draft version.

Möhler, Gregor & Alistair Conkie (1998): Parametric modeling of intonation using vector quantization. In *Proceedings of the Third International Workshop on Speech Synthesis (Jenolan Caves, Australia)*, 311–316.

Peters, Jörg (2005): *Duden – Die Grammatik* Kap. Intonation. Dudenverlag.

R Core Team (2016): *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org/ (letzter Zugriff: 12. 12. 2017).

Rebholz-Schuhmann, Dietrich, Antonio José Jimeno-Yepes, Erik M. van Mulligen, Ning Kang, Jan Kors, David Milward, Peter Corbett, Ekaterina Buyko, Katrin Tomanek, Elena Beisswanger & Udo Hahn (2010), The CALBC silver standard corpus for biomedical named entities – A study in harmonizing the contributions from four independent named entity taggers. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (Hrsg.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta: European Language Resources Association (ELRA).

Reichel, Uwe D. (2014): Linking bottom-up intonation stylization to discourse structure. *Computer Speech & Language* 28(6), 1340–1365. doi: http://dx.doi.org/10.1016/j.csl.2014.03.005. http://www.sciencedirect.com/science/article/pii/S0885230814000242.

Schauffler, Nadja & Schweitzer, K. (2015): Rhythm influences the tonal realisation of focus. In *Proceedings of Interspeech 2015*, Dresden.

Schiehlen, Michael (2003): A cascaded finite-state parser for German. In *Proceedings of EACL 2003*, 163–166. Budapest.

Schweitzer, A. (2010): *Production and perception of prosodic events – evidence from corpus-based experiments*. Dissertation, Universität Stuttgart.

Schweitzer, A. & Ngoc Thang Vu (2016): Cross-gender and cross-dialect tone recognition for Vietnamese. In *Interspeech 2016*, 1064–1068. doi: 10.21437/Interspeech.2016-405. http://dx.doi.org/10.21437/Interspeech.2016-405 (letzter Zugriff: 12. 12. 2017).

Schweitzer, K., Michael Walsh, Sasha Calhoun, Hinrich Schütze, Bernd Möbius, Antje Schweitzer & Grzegorz Dogil (2015): Exploring the relationship between intonation and the lexicon: Evidence for lexicalised storage of intonation. *Speech Communication* 66(0), 65–81. doi: http://dx.doi.org/10.1016/j.specom.2014.09.006. http://www.sciencedirect.com/science/article/pii/S0167639314000727.

Schäfer, Roland (2015): Processing and querying large web corpora with the COW14 architecture. In Piotr Bański, Hanno Biber, Evelyn Breiteneder, Marc Kupietz, Harald Lüngen & Andreas Witt (Hrsg.), *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, UCREL Lancaster: Institut für Deutsche Sprache. http://rolandschaefer.net/?p=749 (letzter Zugriff: 12. 12. 2017).

Seeker, Wolfgang & Jonas Kuhn (2012): Making ellipses explicit in dependency conversion for a german treebank. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (Hrsg.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey: European Language Resources Association (ELRA).

Seeker, Wolfgang & Jonas Kuhn (2013): Morphological and syntactic case in statistical dependency parsing. *Computational Linguistics* 39, 23–55. http://www.aclweb.org/anthology-new/J/J13/J13-1004.pdf (letzter Zugriff: 12. 12. 2017).

Sekine, Satoshi (1997): The domain dependence of parsing. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, 96–102.

Sjölander, Køare & Jonas Beskow (2000): Wavesurfer – an open source speech tool.

Surdeanu, Mihai & Christopher D. Manning (2010): Ensemble models for dependency parsing: Cheap and good? In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 649–652. Los Angeles, California: Association for Computational Linguistics. http://www.aclweb.org/anthology/N10-1091 (letzter Zugriff: 12. 12. 2017).

Syrdal, Ann K., Julia Hirschberg, Julie McGory & Mary Beckman (2001): Automatic tobi prediction and alignment to speed manual labeling of prosody. *Speech Communication* 33(1–2), 135–151. doi: 10.1016/S0167-6393(00)00073-X. http://dx.doi.org/10.1016/S0167-6393(00)00073-X.

Thiele, Gregor, Wolfgang Seeker, Markus Gärtner, Anders Björkelund & Jonas Kuhn (2014): A graphical interface for automatic error mining in corpora. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 57–60. Gothenburg, Sweden: Association for Computational Linguistics. http://www.aclweb.org/anthology/E14-2015 (letzter Zugriff: 12. 12. 2017).

Zarrieß, Sina, Florian Schäfer & Sabine Schulte im Walde (2013): Passives of reflexives: A corpus study. Abstract at Linguistic Evidence – Berlin Special. https://www2.hu-berlin.de/linguistic-evidence-berlin-2013/download/abstracts/Zarriess_Schaefer_Schulte-im-Walde_LinguisticEvidence2013.pdf (letzter Zugriff: 12. 12. 2017).

Alexander Mehler, Wahed Hemati, Rüdiger Gleim und Daniel Baumartz

7 VienNA

Auf dem Weg zu einer Infrastruktur für die verteilte interaktive evolutionäre Verarbeitung natürlicher Sprache

Abstract: In diesem Beitrag untersuchen wir Entwicklungstendenzen von Infrastrukturen in den Digitalen Geisteswissenschaften. Wir argumentieren, dass infolge (1) der Verfügbarkeit von immer mehr Daten über sozial-semiotische Netzwerke, (2) der Methodeninflation in geisteswissenschaftlichen Disziplinen, (3) der zunehmend hybriden Arbeitsteilung zwischen Mensch und Maschine und (4) der explosionsartigen Vermehrung künstlicher Texte ein erheblicher Anpassungsdruck auf die Weiterentwicklung solcher Infrastrukturen entstanden ist. In diesem Zusammenhang beschreiben wir drei Informationssysteme, die sich unter anderem durch die Interaktionsmöglichkeiten unterscheiden, die sie ihren Nutzern bieten, um solchen Herausforderungen zu begegnen. Dabei skizzieren wir mit VienNA eine neuartige Architektur solcher Systeme, welche aufgrund ihrer Flexibilität die Möglichkeit bieten könnte, letztere Herausforderungen zu bewältigen.

Keywords: Architekturen, Forschungswerkzeuge, Interoperabilität, Texttechnologie

1 Einleitung

Infrastrukturen für die Digital Humanities (DH) stehen vor Herausforderungen, die in den kommenden Jahren einen erheblichen Anpassungsdruck auf deren

Alexander Mehler, Goethe-Universität Frankfurt, Robert-Mayer Straße 10, D-60325 Frankfurt a. M., E-Mail: mehler@em.uni-frankfurt.de
Wahed Hemati, Goethe-Universität Frankfurt, Robert-Mayer Straße 10, D-60325 Frankfurt a. M., E-Mail: hemati@em.uni-frankfurt.de
Rüdiger Gleim, Goethe-Universität Frankfurt, Robert-Mayer Straße 10, D-60325 Frankfurt a. M., E-Mail: gleim@em.uni-frankfurt.de
Daniel Baumartz, Goethe-Universität Frankfurt, Robert-Mayer Straße 10, D-60325 Frankfurt a. M., E-Mail: baumartz@stud.uni-frankfurt.de

https://doi.org/10.1515/9783110538663-008

Ausgestaltung ausüben dürften. Diese Erwartung steht in Zusammenhang mit vier Entwicklungen, die nahezu alle sprachbezogenen Wissenschaften betreffen:

1. *Online-Kommunikation:* Andernorts wurde bereits vielfach betont, dass Lesen und Schreiben zunehmend online geschehen, immer öfter auch so, dass die Taktrate ihrer Verzahnung sich der mündlichen Kommunikation annähert (Lobin 2014; Beißwenger & Storrer 2008). Von dieser Tendenz sind genuine Webgenres (Mehler, Sharoff & Santini 2010) ebenso betroffen wie webbasierte Auftritte von zuvor klassischen Printmedien (Lobin 2014). Diese bereits weit vorangeschrittene Entwicklung legt es nahe, in vernetzten, webbasierten, kollaborativ erstellten, diskutierten und oftmals fortwährend überarbeiteten sprachlichen Aggregaten erste Adressen für eine massive Erweiterung oder gar Verschiebung[1] des Objektbegriffs textbezogener Wissenschaften zu sehen.
2. *Sozial-semiotische Vernetzung:* Das Paradebeispiel der Wikipedia (Stegbauer 2009) eröffnet in diesem Zusammenhang den Blick auf eine Datenstruktur, die ein ganzes System nicht nur sprachlicher Ordnungsparameter beobachtbar macht. Dies betrifft neben den Entstehungsgeschichten von Schreibprozessen insbesondere jene Agenten bzw. deren *algorithmischen Identitäten* (Cheney-Lippold 2017), die an letzteren, ihrerseits zunehmend vernetzten Prozessen ursächlich beteiligt sind. Diese Beobachtung legt den Schluss nahe, dass die unter Punkt (1) diagnostizierte Gegenstandserweiterung nicht allein auf die Vernetzung intertextueller bzw. -medialer Aggregate bezogen ist, sondern die wechselseitige Konstitution von Sprechergemeinschaften, Kulturen und sprachlichen Subsystemen (vgl. Everett 2013) einbezieht. Diesen wohl nur im Methodenverbund von Soziologie, Psychologie und Informatik zu fassenden Gegenstand adressieren wir mit dem Begriff des *sozialsemiotischen Mehrebenennetzwerks* (Mehler et al. 2018) und wagen die Prognose, dass solche Netzwerke ins Zentrum informationswissenschaftlicher Arbeit rücken werden.
3. *Methodenzuwachs:* Eine dritte Entwicklung geht von der Informatik aus, betrifft aufgrund ihres Erfolges jedoch alle Bereiche, die um die Teilautomatisierung ihrer Forschungstätigkeit bemüht sind. Es geht um die Fortschritte im Bereich neuronaler Netzwerke, um das so genannte *Deep Learning* also, dem bahnbrechende Leistungen bei der automatischen Segmentierung und Klassifikation zu verdanken sind. Kosko (2017: 497) sieht in der gestiegenen Rechenleistung eine wesentliche Ursache für diesen Qualitätssprung, während Hearst (2017: 339) den Zuwachs an verfügbaren

1 In dem Sinne, dass die Akzeleration der internetbasierten Kommunikation zu einer stetigen Anpassung der informationswissenschaftlichen Aufmerksamkeit führen wird.

Daten und Speicherkapazitäten als weitere Ursachen nennt. Dieser Lesart zufolge erleben wir negativ formuliert keinen algorithmischen Qualitätssprung, sondern werden bloß Nutznießer eines Ressourcenzuwachses. Positiv formuliert jedoch lernen Algorithmen aufgrund dieser Entwicklung immer schneller immer komplexere Muster (Kosko 2017). Dessen ungeachtet dürften ihre Auswirkungen enorm sein, und zwar bezogen auf den Zuwachs an verfügbaren Werkzeugen – ob nun in Form von Webservices oder Komponenten von Entwicklungsumgebungen (z. B. Weka) bzw. Programmiersprachen (z. B. R) –, die immer mehr Aufgaben der automatischen Analyse und Annotation sprachlicher Daten übernehmen. Es entwickelt sich sozusagen ein „Überangebot" leicht handhabbarer Werkzeuge, die die Digitalisierung der jeweiligen Disziplin zu forcieren versprechen, vermeintlich ohne *informatics literacy* aufseiten ihrer Anwender einzufordern. Methodologisch gesprochen entstehen auf diese Weise Blackboxes, und zwar vergleichbar jenen, die bereits für das *Data Mining* im Allgemeinen diagnostiziert wurden (Cheney-Lippold 2017). Während also die unter Punkt (1) und (2) adressierten Entwicklungen einen Forschungsgegenstands-bezogenen Anpassungsdruck erzeugen, geht es hier um die Auswirkungen eines geradezu inflationären Zuwachses an digitalen Methoden auch solcher Disziplinen, welche traditionell der Informatik fernstehen. Gewährleistung von Reproduzierbarkeit (Peng 2011) und algorithmischer Transparenz werden daher zu Schlüsselfunktionen zukünftiger Infrastrukturen.

4. *Hybridisierung:* Für sich genommen mögen letztere Entwicklungen bereits herausfordernd sein. Angesichts einer parallelen Entwicklung, die auf unsere Arbeits- und Kommunikationskultur als Ganzes gerichtet ist, gewinnen sie jedoch erheblich an Bedeutung. Dies betrifft zunächst die fortschreitende Hybridisierung der Arbeitsteilung zwischen menschlichen und künstlichen Agenten im Bereich der allgemeinen Zeichenverarbeitung. Unter textanalytischer Perspektive geht es dabei um die Methodenseite textorientierter Disziplinen. Begleitet wird diese Entwicklung von einer geradezu inflationären[2] Erzeugung künstlicher Texte, deren Autoren in erster Linie vollautomatischen, unabhängig von der zugrundeliegenden Software-Plattform agierenden Software-Agenten[3] bzw. Software Robots (Ferrara et al. 2016: 96) oder schlicht Bots (Geiger 2014: 347) entsprechen und also nur mittelbar jenen Personen zuzurechnen sind, welche die zugrundeliegenden Algorithmen entworfen bzw. implementiert haben. Die

---

2 Zu dieser Einschätzung siehe die Beispiele unten.

3 Geiger (2014) spricht von so genanntem „bespoke code".

hiermit verbundene generative Perspektive betrifft nun die Objektseite textorientierter Disziplinen. Erstere Entwicklung reicht vom Paradigma des *Interactive Machine Learning* (Patel et al. 2010) bzw. des *Human-in-the-loop* (Hoque & Carenini 2015), bei dem Maschinen den Lernprozess implementieren, während Menschen als Innovatoren agieren (siehe Tab. 7.1), bis hin zum *Human Computation* (von Ahn 2008), bei dem Maschinen die Aufgabe der rudimentären Arbeitsorganisation übernehmen (Kosorukoff 2001) und das maschinelle Lernen (wenn überhaupt) nachgeordnet ist. Letztere Entwicklung reicht von *Social Bots* (mit dem Spezialfall der *Conversational Companions* (Wilks et al. 2010) oder *Chatbots*)[4] über *Wiki(pedia) Bots*[5] (Geiger 2014), *News Bots*[6] und *Influence Bots* (Subrahmanian et al. 2016) bis hin zu *Spambots* (Cresci et al. 2017).[7] Am Beispiel von Wikipedia erläutern Geiger & Ribes (2010), inwiefern Bots gar als Agenten eines soziotechnischen Netzwerks aufzufassen sind, das eine Art der verteilten Kognition (Hollan, Hutchins & Kirsh 2000) implementiert, die unter anderem auf die Herausbildung von Qualitätsstandards zielt.

Die eigentliche Bedeutung letzterer Entwicklung betrifft die Erwartung, zukünftig von künstlichen Texten geradezu überschwemmt zu werden, was mehrere Fragen aufwirft:

- Wie nahe ist der Zeitpunkt, ab dem die Wahrscheinlichkeit, einen Text von einem artifiziellen Autor zu lesen, höher sein wird, als einen Text von einem Menschen?
- In welchem Kommunikationsbereich (etwa der Nachrichtenkommunikation (Ferrara et al. 2016; Lokot & Diakopoulos 2016), der Freizeitkommunikation (Ferrara et al. 2016), der Wissenskommunikation (Iosub et al. 2014) oder der Erinnerungskultur (Ferron & Massa 2014)) wird dieser „break-even point“ zuerst erreicht?

---

4 Bot-Software also, die mit menschlichen Agenten in *Online Sozialen Netzwerken* (OSN) „interagieren“, indem sie deren Verhalten simulieren – ob nun zu deren Schaden oder Nutzen (Boshmaf et al. 2012; Abokhodair, Yoo & McDonald 2015; Freitas et al. 2016; Ferrara et al. 2016). Zu dem zugrundeliegenden Interaktionskonzept siehe kritisch Mehler (2010).

5 Die überwiegend Editionsarbeiten und Vandalismusbekämpfung adressieren (Halfaker & Riedl 2012), teils aber auch ganze Artikel (etwa mittels Übersetzung) erzeugen – siehe unten.

6 Die beispielsweise in Twitter *trending topics* ebenso wie Nischenthemen adressieren können, unter anderem mittels Reproduktion, Aggregation oder datenbezogener Ergänzung (Lokot & Diakopoulos 2016; Steiner 2014). Dabei zeigt sich eine Art Skalenfreiheit, wonach sehr wenige Bots einen sehr großen Anteil an gesendeten Tweets bzw. Links produzieren (Larsson & Hallvard 2015).

7 Bots, die rein repetitive Aufgaben der Organisation von Informationssystemen übernehmen, zählen wir ebenso wenig hierzu wie *Retweet Bots* (Gilani et al. 2016).

Tab. 7.1: Kombinationsmöglichkeiten von menschlichen, maschinellen oder gemischten selektierenden oder innovierenden Agenten. Die innere Mensch-Maschine-Matrix (Szenario 1–4) stammt von Kosorukoff (Kosorukoff 2001).

		Selektierender Agent		
		Maschine	**Mensch**	**Gemischt**
Innovierender Agent	Maschine	Szenario 1 *Distributed Genetic Algorithm* (DGA)	Szenario 2 *Interactive Genetic Algorithm* (IGA)	Szenario 3
	Mensch	Szenario 4 *Computer-Aided Design* (CAD)	Szenario 5 *Human-based Genetic Algorithm* (HbGA)	Szenario 6
	Gemischt	Szenario 7	Szenario 8	Szenario 9 *VienNA*

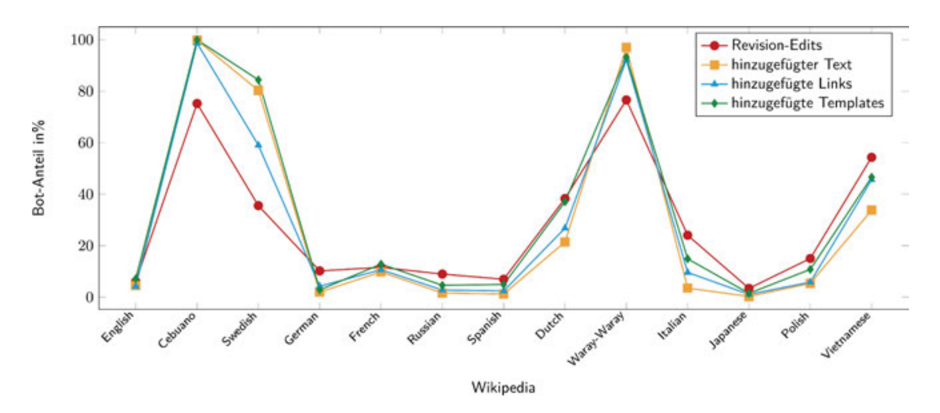


Abb. 7.1: Von Bots autorisierte Anteile an den 13 größten (der Größe nach von links nach rechts geordneten) Wikipedias mit mindestens 1.000.000 Artikeln. Berücksichtigt werden nur Beiträge offiziell registrierter Bots.

– Und was bedeutet diese Entwicklung für die Arbeitsteilung unter den textorientierten Disziplinen? Wird die Informatik hier gar zur dominanten Disziplin, da sie es ist, die Algorithmen zur Erzeugung und Erkennung textgenerierender Bots an erster Stelle erforscht?

Die Wissenskommunikation ist längst in höherem Maße ein Beleg für die mit diesen Fragen umkreiste Entwicklung als es angesichts der für sie einschlägigen Beispiele in Form von Wikipedia und Wiktionary unmittelbar glaubwürdig zu sein scheint. Seien hierzu in Abbildung 7.1 alle Wikipedias betrachtet,

die über mindestens eine Millionen Artikel verfügen. Offenbar stechen zwei Extrembeispiele hervor, und zwar die Wikipedia für *Cebuano* (nunmehr die zweitgrößte Wikipedia) wie auch jene für *Waray-Waray*. In beiden Fällen liegt der Anteil der von Bots autorisierten Texte bzw. Links über 90 % – im Falle von Cebuano werden nahezu 100 % erreicht, durch eine Sprache also, welche eher zu den *low-resource languages* zählt.[8] Mit Vietnamesisch und Schwedisch sind offenbar weitere „Bot-lastige" Wikipedias gegeben. Der Anteilswert von knapp 60 % an Links, den Bots an der schwedischen Wikipedia halten, zeigt z. B., dass die Wahrscheinlichkeit, eine solcherart manifestierte, durch einen Bot erzeugte Textrelation zu rezipieren, den oben erwähnten Break-even-Point bereits überschritten hat. Das bedeutet, dass über intertextuelle Relationen vermittelte, kotextuelle Einbettungen, die maßgeblich Einfluss darauf nehmen, im Kontext welcher Texte ein gegebener enzyklopädischer Artikel oder Textabschnitt rezipiert wird, in solchen Beispielen überwiegend, teils sogar ausschließlich maschinell erzeugt sind. Anders formuliert: Was in welchem textuellen Kontext rezipiert wird, darüber entscheiden immer häufiger Algorithmen und nur noch mittelbar deren Entwickler.[9] Abbildung 7.2 zeigt ergänzend das Beispiel des russischen Wiktionarys, an dem auffällt, dass die mit Abstand aktivsten (an der Kreisunterseite lokalisierten) Agenten abermals Bots sind. In diesem Wiktionary wird lexikalisches Wissen offenbar in hohem Maße durch artifizielle Agenten geformt und vermittelt.

(1) Die Verfügbarkeit von immer mehr Daten über sozial-semiotische Netzwerke, (2) eine Methodeninflation in zuvor rein geisteswissenschaftlichen Disziplinen, die sich methodisch der Informatik annähern, wie auch (3) die zunehmend hybride Arbeitsteilung zwischen Mensch und Maschine und schließlich (4) der Anstieg an artifiziellen Zeichenproduktionen bilden Triebfedern für die Weiterentwicklung von Infrastrukturen, die eine Reihe von informationswissenschaftlichen Anpassungen bzw. Dynamisierungen nach sich ziehen. Um die hiermit verbundenen Diversifikationsmöglichkeiten zu überschauen, knüpfen wir an die Systematik von Kosorukoff (2001) an (siehe Tab. 7.1), die wir um Varianten erweitern, in denen gemischte (künstliche und menschliche) Agenten als Innovatoren (Rekombinatoren) bzw. Selektoren im evolutionstheoretischen Sinne agieren.

---

8 Letzteres Beispiel ist u. a. dadurch erklärbar, dass ein einzelner, nichtmuttersprachlicher Autor (https://ceb.wikipedia.org/wiki/Gumagamit:Lsj) einen Bot namens *Lsjbot* (https://ceb.wikipedia.org/wiki/Gumagamit:Lsjbot (letzter Zugriff 23.10. 2017) verantwortet, mittels dessen er Texte in das Wiki einfügt.

9 Man könnte in solchen Beispielen einen Anlass für das Wiederbeleben traditioneller Enzyklopädien erblicken, insofern diese garantieren können, gänzlich von Menschen verfasst zu sein.

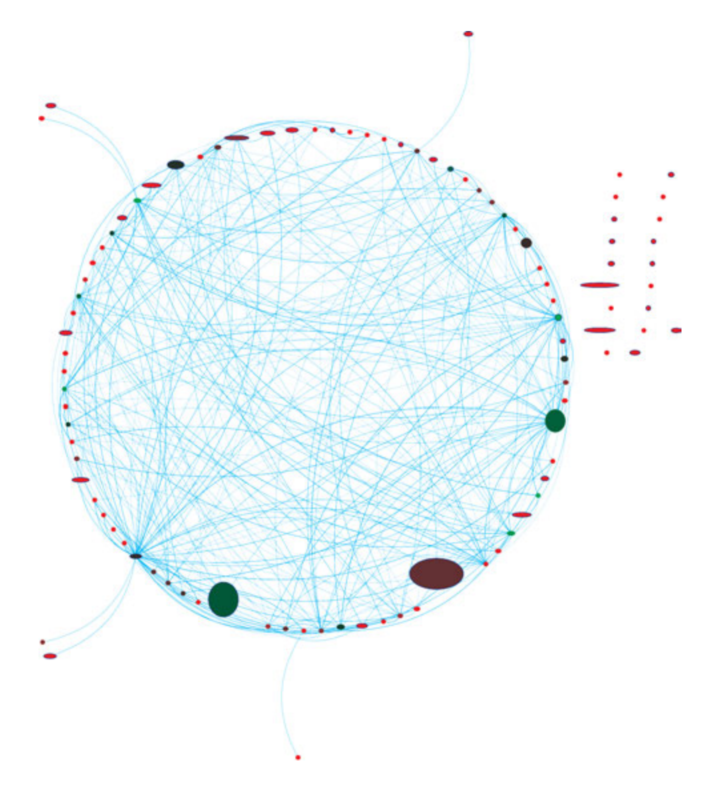


Abb. 7.2: Kollaborationsgraph der 100 aktivsten Agenten des russischen Wiktionarys: Knoten denotieren Agenten, Kanten deren Koautorenschaftsbeziehungen. Die Knotengröße reflektiert den Aktivitätsgrad eines Autors, wobei Knoten von Bots (blau) umrandet sind. Zu den Details der Netzwerk-Berechnung und -Visualisierung, welche auf Brandes et al. (2009) basiert, siehe Mehler et al. (2018).

Gemäß Tabelle 7.1 ist der Standardfall einer Infrastruktur, bei der eine Maschine *M* das jeweilige Inputkorpus analysiert und Menschen die resultierenden Ergebnisse im Zuge ihrer Interpretation evaluieren, eine degenerierte Form eines *Interactive Genetic Algorithm* (IGA), bei dem der Wissenschaftsprozess, in dessen Rahmen es zur Weiterentwicklung bzw. Ersetzung von *M* durch geeignetere Modelle kommt, den entsprechenden Optimierungsprozess implementiert. In Sektion 2 fokussieren wir auf diesen Standardfall und somit auf das gegenwärtig dominante Paradigma von Infrastrukturen.

Indem die Geschwindigkeit von Evaluation und entsprechender Anpassung im Rahmen des zuvor genannten Wissenschaftsprozesses stetig zunimmt, nähern wir uns einem IGA an, bei dem *M* idealerweise selbst zum Gegenstand von Innovationen wird, um immer bessere Analyseresultate aus der Sicht ihrer

menschlichen Interpreten zu erzielen.[10] An dieser Stelle sind Szenarien denkbar, in denen Maschinen den Selektionsprozess unterstützen (Szenario 3) oder umgekehrt Menschen den Innovationsprozess ergänzen (Szenario 8). Beide Szenarien betten Instanzen dessen ein, was Kosorukoff *Human-based Genetic Algorithm* (HbGA) nennt und wofür Wiki-basierte Systeme derzeit die prominentesten Beispiele darstellen (Szenario 5). Dieser Lesart zufolge geht es im Zusammenhang der automatischen Textanalyse bei Szenario 3 und 8 um Ansätze, bei denen Menschen Wikis dazu verwenden, ihre Evaluationen, Korrekturen, Ergänzungen und Interpretationen automatisch erzeugter Analyseresultate zu dokumentieren, so dass diese dazu herangezogen werden können, die zugrundeliegenden Algorithmen zu optimieren. Sektion 3 widmet sich einer Vorstufe dieser Variante von Infrastruktur, und zwar am Beispiel von Wikidition (Mehler et al. 2015).

Die Extremfälle entlang der Hauptdiagonale von Tabelle 7.1, denen zufolge entweder Maschinen (DGA) oder Menschen (HbGA) die Rollen von Innovatoren und Selektoren einnehmen, ruft Szenario 9 auf den Plan, in welchem diese Rollen gemischt besetzt sind. Auf dieses Szenario, dem unserer Auffassung nach die Zukunft wissenschaftlicher Infrastrukturen gehört, fokussiert Beispielgebend Sektion 4. Es geht dabei um eine Architektur von Infrastrukturen für die verteilte interaktive evolutionäre Verarbeitung natürlicher Sprachen, welche auf die optimale Verflechtung von *artificial* und *human computation* zielt. In einem solchen Informationssystem interagieren NLP-Bots als textanalytische Pendants der oben genannten textgenerierenden Bots, und zwar so, dass ihre evolutionäre Weiterentwicklung durch wechselseitige Interaktion (DGA) wie auch durch Interaktion mit ihren menschlichen Nutzern (IGA), die untereinander kommunizieren, um ihre Ziele und Anforderungen weiterzuentwickeln (HbGA), stetig vorangetrieben wird. Szenario 9 zielt folglich darauf, die oben angesprochene Hybridisierung auf die Spitze zu treiben – diesem Szenario dürften zukünftige Infrastrukturen gewidmet sein.

Zusammenfassend gesprochen thematisiert der Beitrag drei aufeinander aufbauende Infrastrukturbeispiele von zunehmender Interaktionskomplexität: Zunächst erläutert Sektion 2 eine Reihe von Bezugsgrößen der Dynamisierung von Infrastrukturen, die von adaptierbaren NLP-Pipelines bis hin zu verteilten Serverarchitekturen reichen. Damit werden zugleich Möglichkeiten und Grenzen des oben genannten Standardfalls benannt. Hierauf aufsetzend thematisiert Sektion 3 die Verzahnung von *Natural Language Processing* (NLP) und Wiki-Prinzip. Zu diesem Zweck wird mit *Wikidition* ein Hybrid-Wiki beschrieben, das

10 Dieses Szenario entspräche – allerdings positiv gewendet – der evolutionären Optimierung von Bots (Cresci et al. 2017).

auf Interaktionsmöglichkeiten im Sinne eines HbGA zielt. Schließlich fokussiert Sektion 4 auf die Aufgabe, sämtliche der zuvor thematisierten Entwicklungstendenzen in einen Anforderungskatalog für Infrastrukturen zu überführen, die auf die optimale Interaktion von Algorithmen und ihren Nutzern zielen. Zu diesem Zweck wird mit VienNA ein Architekturmodell entworfen, das sich dieser Vision – sozusagen mit Blick auf 2020 – annimmt. Sektion 5 fasst die Ergebnisse des Beitrags zusammen und gibt einen Ausblick auf zukünftige Arbeiten.

2 TextImager

Die Verfügbarmachung von immer neuen Methoden für immer mehr Annotationsaufgaben, wie sie in der Einleitung beschrieben wurde, bildet eine Herausforderung für alle bestehenden Infrastrukturen. Folglich bedarf es der Entwicklung flexibilisierter Architekturen, die dieses Wachstum bewältigen können. Dabei ist zu fragen, entlang welcher Dimensionen diese Flexibilisierung zu gewährleisten ist. Zwecks Beantwortung dieser Frage rekurrieren wir auf TextImager (Hemati, Uslu & Mehler 2016; Uslu et al. 2017; Hemati, Uslu & Mehler 2017) als ein Beispiel für Systeme, welche sich an der „klassischen" und also weitgehend linearen, nicht-interaktiven Arbeitsteilung von zuerst algorithmischer (automatischer) Analyse und nachfolgender (menschlicher) Interpretation der Analyseergebnisse orientieren (Hearst 1999).

TextImager[11] stellt NLP-Pipelines für die automatische Analyse textueller Einheiten bereit. Vergleichbar mit Voyant Tools (Rockwell & Sinclair 2016) geht es dabei um die interaktive Visualisierung textueller Strukturen, und zwar so, dass Interaktionen mit Texten zu Anpassungen der mit ihnen verschränkten Visualisierungen führen. TextImager erlaubt daher die Interaktion mit Visualisierungen, um das Browsen in den Inputtexten zu steuern, wie auch umgekehrt das Browsen in Texten, um kontextsensitive Visualisierungen zu erzeugen. Vergleichbar mit DKPro (Eckart de Castilho & Gurevych 2014) nutzt TextImager Apache UIMA, um NLP-Tools zu orchestrieren. Eine der Aufgaben von TextImager besteht darin, die hiermit verbundenen Kombinationsmöglichkeiten zu erweitern. Vergleichbar mit GATE (Cunningham 2002), RapidMiner (Ertek, Tapucu & Arin 2013) und WebLicht (Hinrichs, Hinrichs & Zastrow 2010) wird das Ziel verfolgt, Text-Mining-Werkzeuge auf einer möglichst breiten Basis verfügbar zu machen. Wie bei ConText (Diesner 2014)

11 http://textimager.hucompute.org/ (letzter Zugriff: 30. 4. 2018).

betrifft dies auch netzwerkanalytische Software (Uslu et al. 2017). Im Prinzip soll jedes frei verfügbare Tool für jede natürliche Sprache, das UIMA-konform standardisiert ist, in TextImager integrierbar sein. Zurzeit werden Werkzeuge für Englisch (44), Deutsch (26), Spanisch (22), Französisch (13), Latein (10), Niederländisch (7) , Portugiesisch (8), Chinesisch (7), Italienisch (5), Dänisch (5), Arabisch (4), Türkisch (3) und Bulgarisch (2) bereitgestellt.

Im Folgenden rekurrieren wir auf TextImager, um die oben eingeforderte Dynamisierung entlang von sechs Kriterien zu skizzieren:

1. *Multi-User-Systeme:* Reproduzierbares datenorientiertes Arbeiten erfordert die Möglichkeit einer arbeitsteiligen Organisation unter Kooperationspartnern, die unterschiedliche Rechte an der Analyse und Annotation der zugrundeliegenden Daten halten können. Ausgehend von den Möglichkeiten, die Betriebssysteme zur Bildung entsprechender Arbeitsgruppen bieten, verwendet TextImager hierzu das Rechteverwaltungssystem des *eHumanities Desktops* (Gleim, Mehler & Ernst 2012). Auf diese Weise ist es prinzipiell möglich, Vererbungshierarchien über Entitäten zu definieren, die Subgraphen tripartiter Graphen über Usern (*wer darf*), Funktionen bzw. Werkzeugen (*was* bzw. *womit*) und Daten(-repositorien) (*worauf*) aufspannen, um etwa Rechte an Unter-Arbeitsgruppen einschränkend oder erweiternd zu vererben. Darüber hinaus ist TextImager auch anonym verwendbar und erlaubt somit Nutzungsmodalitäten, wie sie für Wikis typisch sind.
2. *Multi-Service-Systeme:* Automatisierungsfortschritte werden es obsolet machen, dass der Mensch auf allen Ebenen des NLP arbeitsorganisierend agiert, so dass zukünftig eher selbstorganisierende, selbstlernende Maschinen diese Arbeitsprozesse in Gang halten werden (siehe Sektion 4). Dies dürfte insbesondere den Bereich der so genannten Vorverarbeitung (zu der beispielsweise die Lemmatisierung oder das *PoS-Tagging* zählt) betreffen. Es sind aber auch Zusammenschlüsse bestehender Systeme und Entwicklungsumgebungen denkbar, die Arbeitspakete untereinander austauschen, um Spezialisierungsvorteile zu nutzen. Eingedenk dieser Entwicklung bietet TextImager sämtliche seiner Werkzeuge als Webservices an, die folglich Plattform- und Programmiersprachen-unabhängig nutzbar sind. Das mit dieser Vorgehensweise verbundene Dynamisierungsprinzip bezieht sich folglich darauf, Infrastrukturen unabhängig von HCIs nutzbar zu machen.
3. *Multi-Pipeline-Systeme:* Infolge der oben diagnostizierten Methodenvielfalt stehen für dieselbe Aufgabe (z. B. *Tagging*) immer mehr Algorithmen (z. B. *Conditional Random Fields*) und Werkzeuge (z. B. *MarMoT*) bereit, die ihrerseits Parameterräume aufspannen und somit Entscheidungsalternativen auf zumindest drei Ebenen erzeugen. Zudem hat sich im Zuge der Entwicklung der Computerlinguistik als wissenschaftliche Disziplin ein rudimen-

täres Prozessmodell etabliert, das eingeübte Abfolgeregularitäten solcher Aufgaben festlegt. So setzt etwa das *Semantic Role Labeling* vielfach auf Parsing auf, das wiederum Ergebnisse von Lemmatisierung und PoS-Tagging voraussetzt. Ein solches Prozessmodell erlaubt Abstraktionen auf zumindest vier Ebenen: im Hinblick auf (a) die Klassifikation von Teilaufgaben zum Zwecke der Methodenübertragung (PoS-Tagging und NER als Beispiele für *Sequence Labeling*), (b) die variierende Instanziierung des Prozessmodells zwecks Fokussierung auf Teilaufgaben, (c) die Orchestrierung von Teilmengen von Werkzeugen für dieselbe Aufgabe zur Erzeugung von Skaleneffekten (etwa mittels *majority voting*) und schließlich (d) die Neuordnung von Prozessschritten zwecks Erzielung von Seiteneffekten (wenn etwa Parsing-Ergebnisse dazu verwendet werden sollen, das PoS-Tagging nachzubessern). In diesem Zusammenhang bietet TextImager aufgrund seiner UIMA-basierten Architektur die Möglichkeit, seine Dienste mittels alternativer Pipelines zu ordnen, die zur Laufzeit erstellt werden. Auf diese Weise werden zumindest die Dynamisierungsschritte (b) und (c) adressiert.

4. *Multi-Server-Systeme:* Zuwachs an Ressourcen bedeutet auch, dass immer mehr Rechnerkapazität benötigt wird, um die Arbeit von NLP-Tools auf immer größeren Datenmengen zu orchestrieren. TextImager begegnet dieser Herausforderung durch Verteilung seiner Services auf im Prinzip unbegrenzt viele Server, und zwar so, dass eine $n:m$-Relation von Diensten und Servern entsteht. Dieser Ansatz erlaubt das Ein- und Ausschalten bzw. Readressieren immer neuer Server und Services zur Laufzeit. Infolge dieser Dynamisierung muss keiner der Server mehr zusammen mit dem orchestrierenden System gehostet werden. Infrastrukturen dieser Art können folglich als Schnittstellen zu Multi-Server-Systemen ausgebaut werden, die aufgrund dezentraler Initiativen und Weiterentwicklungen stetig wachsen, ohne das eingangs erwähnte Kapazitätsproblem aufzuwerfen. Das ermöglicht es schließlich, dass Nutzer eigene Dienste, die auf ihren Servern liegen, mit Diensten etwa von TextImager kommunizieren lassen, um die verteilte Erledigung ihrer Annotationsaufgaben mittels dieser Infrastruktur zu verwalten. Dabei geht es auch um nicht frei zugängliche Ressourcen (Korpora, Tools etc.), die auf proprietären Servern verbleiben, während sie im Verbund mit frei zugänglichen Ressourcen orchestriert werden. Darüber hinaus adressiert dieser Ansatz die Verarbeitung von *Big Data*, indem er die verteilte Verarbeitung gleichzeitiger Anfragen einer Vielzahl von Usern ermöglicht.
5. *Multi-Datenbank-Systeme:* Ein Spiegelbild des Methodenzuwachses für immer mehr Annotationsaufgaben besteht in der Diversität resultierender

Repräsentationen ob nun in Form dokumentorientierter (XML-basierter) Baumstrukturen, datenorientierter Graphstrukturen oder numerischer Verteilungen. Um auch in diesem Bereich Skalengewinne zu erzielen, bedarf es eines Multi-Datenbank-Systems, das mehrere Datenbankmanagementsysteme (DBMS), welche für die Eigenheiten solcher Repräsentationen optimiert sind, verwalten kann. So eignen sich beispielsweise Graphdatenbanken wie Neo4j für die Repräsentation simpler gerichteter Graphen, während MongoDB für die Verwaltung dokumentorientierter Strukturen besser ausgerüstet ist. Blazegraph wiederum eignet sich für die Verwaltung RDF-basierter Daten, wofür das Wikidata-Projekt ein anwendungstechnisches Aushängeschild ist. TextImager adressiert auch diesen Aspekt der Dynamisierung. Leitbild hierfür ist jene Offenheit, die bereits Multi-Server-Systeme charakterisiert: Es geht darum, immer neue Datenbanken in das System integrieren zu können, um noch immer bestehende Repräsentationsschranken – etwa im Hinblick auf die Analyse von Hypergraphen – angehen zu können. Die dahinter stehende Erwartung betrifft die Erschließung von Skaleneffekten, welche aus der gleichzeitigen Analyse unterschiedlicher Daten und Annotationen derselben sprachlichen Einheiten resultieren. Dahinter steht die Prognose, dass künftige NLP-Infrastrukturen in diesem Sinne hochgradig multimodal sein werden.

6. *Verzahnung:* Indem sie „*Machine Learning*-lastiger“ werden, transformieren sich die Digital Humanities sozusagen zu Computational Humanities (Biemann et al. 2014). Je herausfordernder jedoch die geisteswissenschaftliche Interpretationsarbeit, desto größer der Aufwand der „manuellen“ Annotation von Trainings- und Testdaten, die dazu benötigt werden, *Machine Learning*-Verfahren für die Annotation von Textsegmenten zu trainieren, auf die die anvisierten Interpretationen bezogen oder potentiell beziehbar sind. Diese Aufgabe erfordert eine enge Verzahnung von automatischer Analyse und manueller Annotation. Es geht darum, letztere Prozesskette dahingehend zu dynamisieren, dass der Prozess der manuellen Annotation durch das NLP massiv unterstützt wird, etwa durch Recommender-Systeme zur fortlaufenden Generierung von Annotations- und Vervollständigungsvorschlägen, die der Annotator idealerweise nur auszuwählen braucht, um seine Arbeit zu erledigen. Eingedenk dieser Dynamisierungsanforderung integriert TextImager Werkzeuge für die Annotation von rhetorischen und propositionalen Strukturen, um die Verschmelzung von *Human Computation* und *Machine Learning* voranzutreiben. Dabei geht es letztlich um die Aufhebung der statischen Abfolge von manueller Annotation, automatischer Analyse und nachfolgender Interpretation.

Diese sechs Bezugsgrößen skizzieren Entwicklungsperspektiven von Infrastrukturen, die bei aller Unabdingbarkeit noch immer auf expertenseitige Nutzer fokussieren und damit eine *informatics literacy* einfordern, wie sie von der Mehrzahl ihrer Nutzer noch lange nicht (wenn überhaupt jemals) erbracht (werden) wird. Infolge der linearen Ordnung von automatischer Analyse und intellektueller Interpretation lassen solche Systeme zudem jene Form von artifizieller Interaktivität zwischen Menschen und Maschinen vermissen, die auf die stetige Anpassung und Optimierung der involvierten Maschinen gerichtet ist (Mehler 2010). Klassische Systeme wie Voyant Tools oder TextImager sind genauer insofern nicht-interaktiv, als die Art und Weise der Interaktionen, die sie implementieren, nicht die Dispositionen der eingesetzten Algorithmen für zukünftige Interaktionen mit ihren menschlichen Nutzern oder gar untereinander verändert (Mehler 2010). Anders ausgedrückt: Die Verwendung von TextImager setzt keine Lernprozesse aufseiten seiner Algorithmen in Gang. Folglich bedarf es – ganz im Sinne von Kriterium (6) – eines Systems, das sehr viel stärker die Interaktionsmöglichkeiten mit menschlichen Interpreten automatisch erstellter „Interpretations-Vorprodukte" in den Vordergrund rückt und dabei gleichzeitig der Gefahr methodischer Blackboxes (siehe Sektion 1) begegnet, indem es seinen Rezipienten weitreichende Interaktionsmöglichkeiten bietet, um letztere Informationsangebote zu diskutieren, zu adaptieren und gegebenenfalls zu optimieren. Dies ist die Aufgabe von Wikidition.

3 Wikidition

Oberhalb der Ebene von Einzeltexten geht es bei linguistischen Netzwerken darum, die kohärenzstiftenden Verknüpfungsrelationen ganzer Systeme solcher Einheiten horizontal (d. h. von Einheiten derselben Ebene) wie auch vertikal (d. h. von Einheiten verschiedener Ebenen) sichtbar zu machen, um schließlich Ausblicke auf Prozesse der Sprachevolution im Kontext sozialsemiotischer Netzwerke (siehe Sektion 1) zu gewinnen. Angesichts des in Sektion 1 diagnostizierten Methodenzuwachses ist dabei zu fragen, wie immer neue Methoden nutzbar werden sollen, um solche Vernetzungsrelationen zu identifizieren und zu annotieren, und zwar so, dass Rezipienten die Annotationsergebnisse interaktiv weiterverarbeiten können. Zur Beantwortung dieser Frage rekurrieren wir auf Wikidition (Mehler et al. 2015) als Beispiel für eine Infrastruktur, die die automatische Annotation mit dem Wiki-Prinzip (Leuf & Cunningham 2001) verbindet und folglich eine bei weitem stärkere Verflechtung von *artificial* und *human computation* realisiert, als sie mit TextImager und verwandten Systemen derzeit gegeben ist (siehe Sektion 2). Die Kernfunk-

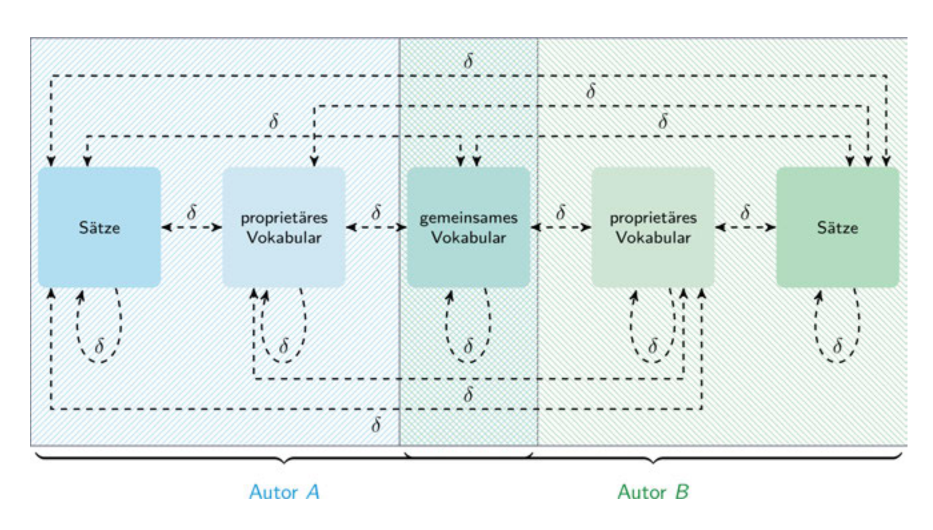


Abb. 7.3: Polypartite Mehrebenen-Netzwerke als das Ergebnis von Wikidition.

tionen von Wikidition lassen sich wie folgt zusammenfassen (zu den Details siehe Mehler et al. 2015):

1. *Linkifizierung:* Ausgehend vom jeweiligen Inputkorpus (Mehler, Wagner & Gleim (2016) unterscheiden neun Varianten solcher Korpora) wird jedes seiner Segmente auf Wort-, Satz- und Textebene durch einen separaten Wiki-Artikel repräsentiert, seine syntagmatischen Kontiguitäts- und paradigmatischen Similaritätsrelationen exploriert und schließlich mittels Hyperlinks manifestiert. Auf diese Weise entsteht ein Mehrebenennetzwerk, in dem jedes Token mit seinem Type und jeder Type mit allen seinen Token verlinkt ist, so dass das resultierende Netzwerk im Sinne kommutierender Diagramme aus jederlei Perspektive (der Wort-, Satz- oder Textebene) in jederlei Richtung traversierbar wird.
2. *Lexikonisierung:* Voraussetzung für die Linkifizierung ist die automatische Extraktion eines Korpus-spezifischen Lexikons, das Wikidition in Form eines eingebetteten Wiktionarys bereithält. Auf diese Weise werden autorenspezifische Eigenheiten der im Inputkorpus enthaltenen Texte zugänglich. Dies bringt Abbildung 7.3 zum Ausdruck, bei der das *gemeinsame Vokabular* auf mehrere Sätze verlinkt, die von mindestens zwei Autoren stammen, während *proprietäres Vokabular* je auf Sätze eines Autors bezogen ist.
3. *Interaktivität:* Da Wikidition im technologischen Sinne ein MediaWiki darstellt, das um Visualisierungsmethoden von TextImager erweitert ist, sind alle seine Inhalte durch alle dazu autorisierten Nutzer änderbar und erwei-

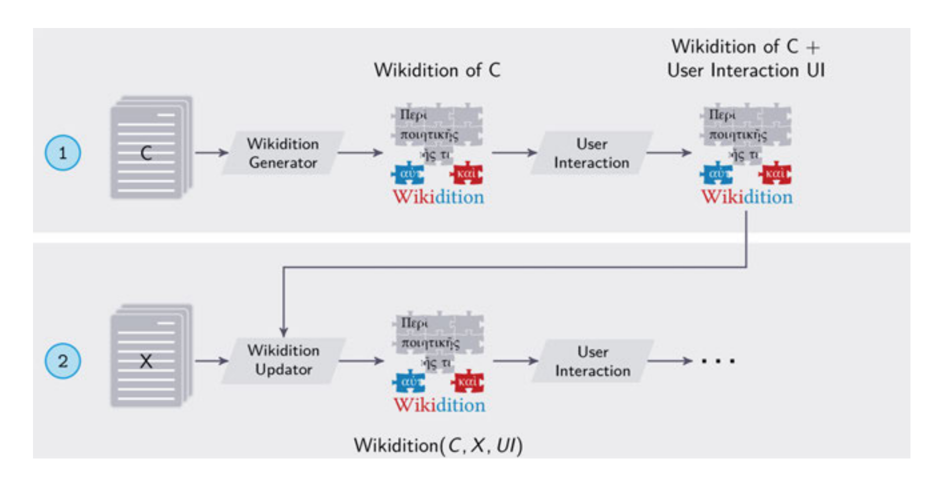


Abb. 7.4: Zur Verschränkung von *artificial* und *human computation* in Wikidition ausgehend von Korpus *C* und seiner Ergänzung um Korpus *X* (vgl. Mehler et al. 2015).

terbar. Das hat den Vorteil, dass alle automatisch erzeugten Annotationen Revisionen unterzogen werden können, und zwar so, dass diese Revisionen nach dem Wiki-Prinzip interaktiv optimiert werden können. Wikidition ermöglicht folglich die Implementation eines *Human-based Genetic Algorithm* im Sinne von Kosorukoff (2001) (siehe Tab. 7.1). Anders als der nachfolgend vorzustellende Ansatz wirkt sich die hierdurch implementierte Interaktivität jedoch nicht auf die fortschreitende Anpassung der involvierten NLP-Routinen aus: Wikidition ermöglicht derzeit daher keine artifizielle Interaktivität im Sinne von Mehler (2010).

4. *Reproduzierbarkeit:* Ergänzt werden die Kernfunktionen durch die gleichzeitige Verfügbarmachung jener Modelle, die zwecks Generierung der Wikidition berechnet wurden. Auf diese Weise werden die betroffenen Modelle auch außerhalb der jeweiligen Wikidition nachnutzbar und reproduzierbar.

Den Ablauf der Erstellung einer Wikidition zeigt Abbildung 7.4: Im ersten Schritt generiert der Wikidition-Algorithmus das oben genannte Mehrebenennetzwerk, das durch nutzerseitige Interaktionen revidiert und erweitert werden kann, ehe Korpora und daraus extrahierte Lexika ergänzt und mit den vorhandenen Ressourcen vernetzt werden. Wikidition ist – wie schon TextImager – sequenziell organisiert: Auf die initiale maschinelle Verarbeitung seiner Textbasis folgt ihre Überarbeitung durch menschliche Agenten, die ergänzen oder korrigieren können, was die maschinelle Verarbeitung unvollständig oder fehlerhaft annotiert hat. Nachfolgend operierende maschinelle Agenten erweitern vorrangig die Textbasis, ohne zuvor erzeugte Annotationen oder ihre eigene

Lernbasis einer maschinellen Revision zu unterziehen. Anders ausgedrückt: Wikidition zielt nicht auf die Optimierung der involvierten NLP-Routinen, sondern auf die andauernde Annotation und Vernetzung der thematisierten Sprachressourcen. Verfahrensinnovationen erfolgen extrinsisch durch Weiterentwicklung von TextImager, auf dem Wikidition basiert. Die verwendeten NLP-Routinen bilden somit einen prozeduralen Input, der mittels einer vorgegebenen Pipeline orchestriert wird. Aus dem Blickwinkel zukünftiger Texttechnologien bietet dieses Szenario eine Reihe von Ansatzpunkten für eine Weiterentwicklung hin zu Verfahren der *verteilten evolutionären Verarbeitung natürlicher Sprachen*, die vor allem auch auf die Optimierung der involvierten NLP-Routinen zielt. Dies wird nachfolgend skizziert.

4 VienNA

Die Grundlage des nun zu skizzierenden Textanalysemodells bildet das Konzept des *Human-based Evolutionary Computing* (HbEC) (Nickerson 2013), das auf menschliche Agenten rekurriert, um Prozessschritte des zugrundeliegenden Evolutionären Algorithmus (EA) zu implementieren bzw. mittels *Human Computation* (HC) (Michelucci 2013) zu ergänzen. Der Vorteil von HbEC besteht in der Integration einer memetischen Komponente (Nickerson 2013), mittels derer die Selektion von Genotypen durch Rekurs auf den Wissens- bzw. Interessenshintergrund von Nutzern gesteuert werden kann. Die Implementation dieser Komponente mittels des für Wikidition konstitutiven Wiki-Prinzips und folglich die Verteilung des HC auf ein Netzwerk interagierender Prosumenten (Tapscott & Williams 2008) – im Folgenden spezieller Wikiditoren genannt – bedeutet dann, dass letzteres Wissen als das Ergebnis eines Wiki-vermittelten Koordinationsprozesses resultiert. Darüber hinaus soll der Koordinationsprozess auf die beteiligten NLP-Komponenten ausgedehnt werden, so dass diese als „eigenständige" Koordinationspartner untereinander ebenso wie mit den beteiligten Prosumenten interagieren. Auf diese Weise soll ein EA implementierbar werden, der Szenario 9 aus Tabelle 7.1 adressiert, in dessen Rahmen alle NLP-Komponenten zu *NLP-Bots* mutieren. Dieser Modellrahmen ist entlang folgender Bezugsgrößen auszuarbeiten:[12]

1. *Integration von Evolutionary und Social Computing*: An erster Stelle geht es um die bereits thematisierte Weiterentwicklung des HbEC zu einem *Distri-*

12 Dabei sprechen wir von Aktualgenese als einem Prozess auf einer kurzfristigeren Zeitskala, und zwar im Verhältnis zur jeweiligen Ontogenese.

buted Human-based Evolutionary Computing (Michelucci 2013) bzw. *Social Computing* (Michelucci 2013), wie es für Wikidition konstitutiv ist.

2. *Ontogenese von NLP-Bots:* An zweiter Stelle geht es um die Dynamisierung der involvierten NLP-Komponenten, die nicht länger einen zwar prozeduralen, aber statischen, im aktualgenetischen Sinne also konstanten Input bilden, sondern selbst einer Evolution unterliegen sollen.
3. *Aktualgenese von Annotationen:* Basierend auf den vorangehenden Dynamisierungsschritten sollen die anvisierten Annotationen der zugrundeliegenden Ressourcen (Texte, Lexika etc.) fortlaufend erzeugt bzw. angepasst werden, und zwar mit dem Ziel ihrer fortgesetzten Optimierung unter Bedingungen, die aus dem Verhalten der beteiligten Prosumenten bzw. kooperierenden oder konkurrierenden Bots resultieren. Annotationsprozesse werden in vergleichsweise kurzen Zeiträumen vollzogen, so dass bei jedesmaliger Verwendung der betroffenen Wikidition deren stetige Veränderung beobachtbar wird.
4. *Genese der Repräsentationen sprachlicher Einheiten:* In Bezug auf letztere Aktualgenese unterscheiden wir zwei Prozessebenen der Repräsentation sprachlicher Einheiten: zum einen die Genese von Merkmalen bzw. Zähleinheiten zu annotierender Objekte und zum anderen die Genese von Annotationseinheiten (als Manifestationen der anvisierten Interpretationen[13]). Man beachte, dass Annotationseinheiten aus der Sicht bestimmter Bots Zähleinheiten aus der Sicht anderer Bots sein können. Um die hiermit verbundene Offenheit des operativen Textrepräsentationsmodells (Mehler 2007) beherrschbar zu machen, gehen wir von einer Grammatik zur Generierung von Datentypen im Rahmen eines verallgemeinerten Klassifikationsszenarios aus (siehe unten).
5. *Genese des Agentennetzwerks:* Die NLP-Bots sind schließlich zu soziotechnischen Gemeinschaften zu integrieren, in denen sie mit ihresgleichen wie auch mit Prosumenten interagieren, und zwar so, dass ihre Orchestrierung selbst zum Gegenstand eines evolutionären Prozesses wird. Auf dieser Grundlage ist zu vermeiden, dass die Orchestrierung etwa in Form einer statischen NLP-Pipeline vorgegeben wird. Das Agentennetzwerk repräsentiert, welcher menschliche oder künstliche Agent mit welchem anderen Agenten auf welche Weise (etwa Daten liefernd oder nachnutzend) interagiert. Das Netzwerk bildet eine emergente Eigenschaft von VienNA, die auf einer Zeitskala zwischen Aktual- und Ontogenese fluktuiert.

---

13 Im einfachsten Fall der Textklassifikation sind dies Klassenlabels.

6. *Phylogenese von NLP-Bots:* Auf der äußersten (langfristigsten) Zeitskala sind schließlich jene Evolutionsprozesse anzusiedeln, die die Herausbildung von Klassen von NLP-Bots und deren „Abstammungsverhältnisse“ betreffen. Dabei handelt es sich um Prozesse, die derzeit noch immer vollständig in Menschenhand liegen, etwa dann, wenn es darum geht, z. B. das Lemmatisieren, PoS-Tagging, *Dependency Parsing* oder das *Argument Mining* als verschiedene, teils interdependente NLP-Aufgaben abzugrenzen und entsprechende Algorithmen zu entwickeln. Aus konzeptioneller Sicht versteht sich unser Ansatz als eine Architektur, in der schließlich auch diese Prozesse zum Gegenstand eines evolutionären Prozesses werden sollen.

Eine NLP-Architektur, die zumindest die Komponenten (1–5) beinhaltet, und also die Ontogenese von Bots und eine hierauf beruhende Aktualgenese von Textrepräsentationen vorsieht, bezeichnen wir als *v*erteilte evolut*i*onär*e* *n*euronalplastische *N*LP-*A*rchitektur (VienNA). Unter Hinzuziehung eines Modells der Phylogenese von Bots thematisiert VienNA folglich mehrere Zeitskalen, deren Dynamiken zum gegenwärtigen Stand des NLP mit Ausnahmen im Bereich der Aktualgenese und teils der Ontogenese noch immer fest in Menschenhand liegen.

Eine fundamentale Herausforderung der Umsetzung von VienNA bildet die Begrenzung des Textrepräsentations- bzw. Annotationsraums, den die beteiligten NLP-Bots infolge ihrer andauernden Arbeiten aufspannen. Es geht dabei unter anderem darum, zu verhindern, dass die Bots diesen Raum stetig vergrößern – auch im Hinblick auf Annotationen, die die involvierten Wikiditoren positiv evaluieren. Diese Problematik kann am Beispiel der Wikidition aus Sektion 3 erläutert werden, deren lexikalische Einheiten im Schnitt über hunderte Links mit Einheiten derselben oder anderer Sprachebenen verknüpft werden. Würden Bots in einem solchen Szenario weitgehend ungehindert intra- und interrelationale Relationen annotieren, erreichte deren Zahl schnell eine Größe, die die zugrundeliegende DB nicht mehr erfassen kann und aus der Sicht der Prosumenten unzumutbar, da nicht rezipierbar wäre. Folglich ist in VienNA die Selektion von Annotations-Genotypen an eine oder mehrere Fitness-Funktionen zu binden, die unter anderem die Raumkomplexität, die Rezipierbarkeit sowie die Filter- und Kontextualisierbarkeit von Genotypen ebenso reflektiert wie deren Prosumenten-orientierte Interpretierbarkeit, Nützlichkeit[14] oder deren Innovationsgrad. Man kann sich die entsprechen-

14 Was interpretierbar ist, ist nicht notwendigerweise auch nützlich.

den Fitness-Funktionen als ein System vorstellen, das einen *Constraint Satisfaction Process* (CSP) induziert: Erzeugt ein Bot einen Annotationsvorschlag (möglicherweise infolge seiner evolutionären Weiterentwicklung), so hat sich dieser etwa unter Beachtung der für den CSP geltenden Raumkomplexitätsschranke dahingehend zu bewähren, dass er fitter ist als mindestens eine der bereits bestehenden Annotationen, deren Speicherbedarf ersetzungshalber zu beanspruchen wäre. Hierzu kann die Fitnessfunktion auf die sich wandelnden Interessen der Wikiditoren rekurrieren, um eine Erstarrung der Wikidition zu verhindern: fitter ist sozusagen das, was die je aktuellen, prinzipiell aber wechselnden Interessen der Prosumenten besser bedient.

Um VienNA als Instanz eines EA zu konstruieren, der zumindest auf den zwei genannten Zeitskalen operiert, orientieren wir uns an der Wiedergabe eines EA nach Nickerson (2013). Dieser beinhaltet zunächst genau eine Zeitskala evolutionärer Dynamik (siehe Algorithmus 1). Ziel ist es, die Verfahrensschritte eines solchen EA im Sinne der anvisierten Dynamisierung zu konkretisieren. Hierzu ist anzugeben, wie sich die Zeitskalen von VienNA in Algorithmus 1 einbetten lassen. Um angesichts der Vielfalt von NLP-Aufgaben die Machbarkeit der Darstellung zu wahren, konzentrieren wir uns beispielgebend auf zwei Klassen von Aufgaben: Textkategorisierung (Joachims 2002) und Textähnlichkeitsmessung (Bär et al. 2012). Wir beginnen mit der Textkategorisierung, mit deren Hilfe einstellige Relationen gelernt werden.

Algorithmus 1: Schema eines *Evolutionären Algorithmus* (EA) nach Nickerson (2013).

```
01 erzeuge eine initiale Generation G von Lösungen;
02 evaluiere G;
03 while G lässt sich verbessern do
04 |  erzeuge eine Menge K von möglichst vielen Paaren von Elementen
   |     aus G;
05 |  foreach Paar aus K do
06 |  |  erzeuge Nachkommen mittels Kombination dieser Paare;
07 |  |  mutiere die Nachkommen;
08 |  |  evaluiere die Nachkommen;
09 |  end
10 |  selektiere die nächste Generation aus der Menge der Eltern und
   |     Nachkommen;
11 end
12 return Menge der besten Lösungen;
```

4.1 Einstellige Relationen

Textkategorisierung ist als Konkatenation

$$g \circ f : C \to \mathscr{L} \tag{1}$$

zweier Funktionen aufzufassen, bei der $f : C \to X$ jedem Text(segment) $t \in C$ des Korpus C eine Textrepräsentation (etwa einen Vektor im Sinne des Vektorraummodells) als Element der Repräsentationsmenge X zuordnet, die die Funktion $g : X \to \mathscr{L}$ auf Klassenlabels der Menge $\mathscr{L}$ abbildet. Man beachte, dass die Konstituenten der Repräsentation $f(t)$ nicht t entstammen müssen.[15] f und g werden i. d. R. getrennt implementiert. Beide Funktionen besitzen daher unterschiedliche Parameterräume, die nachfolgend separiert betrachtet werden.

– *Zum Parameterraum von g:* Wir gehen zunächst davon aus, dass g als SVM implementiert wird. Der g-induzierte Parameterraum beinhaltet dann Klassen von Kernels (z. B. *linear*, *polynomial*, ..., *string kernel*, *tree kernel*, *graph kernel*), die jeweils Kernel-spezifische Parameter induzieren. Ein solcher Parameterraum lässt sich unmittelbar zum Gegenstand eines EA machen, dessen Lösungsraum aus Binärvektoren besteht, und zwar indem die Kombination von Lösungspaaren (Algorithmus 1, Zeile 6) ebenso wie deren Mutation (Zeile 7) mittels Bitoperationen implementiert wird, wobei Parameter und Dimensionen solcher Vektoren eineindeutig aufeinander abgebildet sind. Demgegenüber gestaltet sich die Suche in den Räumen jener Parameter als schwieriger, die die Gestalt von g festlegen. Am Beispiel mehrschichtiger Feedforward-Netze lässt sich ein solcher Raum als mehrdimensionales Grid auffassen, in dem zum Zwecke der Optimierung ein *Random Walk* durchgeführt wird. Im zweidimensionalen Fall lassen sich so etwa Kombinationsmöglichkeiten der Parameter *i-ter Layer* und *Anzahl Neuronen* abbilden. Mit Hilfe solcher Zufallsbewegungen lassen sich Rekombinationen und Mutationen im Sinne von Zeile 6 bzw. 7 aus Algorithmus 1 abbilden. Wir unterteilen folglich den Parameterraum von g in den Raum R_1 jener Parameter, die die Klasse von Funktionen (linearer Kernel vs. String-Kernel etc.), denen g angehört, festlegen, und den Raum R_2 jener Parameter, die den Phänotyp von g als Instanz dieser Klasse bestimmen. Rekombinationen und Mutationen in R_1 bedingen dann solche in R_2 – an dieser Stelle fehlt der Platz, dies ausführlich darzustellen.

15 Etwa bei der Merkmalsexpansion in Bezug auf ein Lexikon. Alternativ, wenn t ein Segment eines Texts t' ist, kann $f(t)$ Merkmale aus der Umgebung von t in t' beinhalten.

- *Zum Parameterraum von f:* Aufgabe der Parameteroptimierung ist nun die Erzeugung von Merkmalen zur Charakterisierung von Einheiten aus C. Wir gehen davon aus, dass diese Merkmale mittels einer Grammatik rekursiv aufzählbar sind. Da die Spezifikation einer solchen Grammatik die Grenzen des Beitrags sprengte, beschränken wir uns auf einen skizzenhaften Modellausschnitt. Unser Ansatz besteht darin, (1) Textsegmente als Mengen, Pfade, Bäume oder allgemeinere Graphen zu repräsentieren, um (2) hieraus häufige Teilmengen (z. B. *frequent itemsets*, Aggarwal & Han 2014), Teilsequenzen (z. B. *k*-Skip-*n*-Gramme, Guthrie et al. 2006), Teilbäume (Zaki 2005) oder Teilgraphen (z. B. *motifs*, Milo et al. 2002) zu extrahieren, so dass (3) diese im Folgenden als Motive bezeichneten Muster ihrerseits in ihrer Mengen, Pfad-, Baum- oder Graph-basierten Zusammenhangsstruktur untersuchbar werden, und zwar mittels aggregierter Motive.[16] Die Motive, die auf der jeweiligen Rekursionsstufe extrahiert werden, bilden Merkmale in Form von Ereignissen, die sich aus zugrundeliegenden Zähleinheiten (wie etwa Lexeme) zusammensetzen. Hieraus resultieren Häufigkeitsverteilungen der Motive über den Instanzen des beobachteten Segmenttyps (Merkmalsträger erster Ordnung wie z. B. Sätze), die mit Hilfe von Funktionen (wie Lage, Streuungs-, Konzentrations- oder Gestaltparameter) schließlich auf die Ausprägungen entsprechender Merkmale von Einheiten $x \in C$ des Zieltyps (Merkmalsträger zweiter Ordnung wie z. B. Texte) abgebildet werden können. Dabei müssen Merkmalsträger zweiter Ordnung nicht jene erster Ordnung enthalten[17] und können sogar mit diesen zusammenfallen. In diesem Ansatz spannen unter anderem die Auswahlgesamtheit für Merkmalsträger erster Ordnung, die Länge bzw. Größe von Motiven, Signifikanzschranken für deren Häufigkeit, die Rekursionstiefe zur Berechnung aggregierter Motive sowie Funktionen zur Charakterisierung von Häufigkeitsverteilungen von Merkmalen einen Parameterraum auf, der wieder zum Gegenstand eines *Random Walks* gemacht werden kann, wobei für jede solcherart „aktivierte“ Parameterkonstellation die Merkmalsausprägungen der Einheiten aus C zu berechnen sind. Nachfolgend denotieren wir diesen Raum mit R_3. Offenbar ist R_3 derart groß und letztere Berechnungen dermaßen komplex, dass eine Zufallsbewegung, die mit dem ein-

16 Dabei bilden Pfade etwa Tokenfolgen ab, während Bäume zur Abbildung von Dependenzstrukturen und ((un-)gerichtete) Graphen zur Abbildung von Assoziationsnetzwerken sprachlicher Einheiten (z. B. der Wortebene) dienen. Aggregationen von Motiven werden beispielsweise in Mehler (2006) beschrieben.

17 Im umgekehrten Fall werden Merkmale aus der Umgebung eines Segments auf dieses abgebildet. Dies geschieht beispielsweise bei der Lemmatisierung.

fachsten Setting in Form eines *Bag-of-elementary-Features*-Modell) beginnt, die einzig effiziente Möglichkeit seiner Traversierung sein dürfte.

Dieser Ansatz deckt eine recht allgemeine Klasse von NLP-Aufgaben ab, wie z. B. Lemmatisierung, PoS-Tagging, Sentimentanalyse oder Textkategorisierung. Er beinhaltet das klassische *Bag-of-words*-Modell ebenso wie Modelle basierend auf Grammen höherer Ordnung oder auf Mustern in Form von Teilgraphen. Seine Grundidee besteht darin, jedwedes Muster (mengenmäßiger, sequenzieller oder graphenmäßiger Art) auf numerische Merkmalsvektoren abzubilden, die zum Input von Neuronalen Netzwerken ebenso gemacht werden können wie von SVMs. An dieser Stelle stellt sich die Frage, warum man nicht einfach auf Neuronale Netze rekurriert, um das hier skizzierte Merkmalsauswahlproblem zu lösen. Der Grund besteht darin, dass unser Ansatz insofern transparenter ist, als er darauf zielt, Merkmale mittels Bitoperationen explizit „ein- oder auszuschalten“, so dass man Lösungen als Ergebnisse entsprechender Optimierungsdurchläufe unmittelbar ablesen kann, welche Merkmalskonstellationen für sie ausschlaggebend waren.

4.2 Binäre Relationen

Anders als die Kategorisierung ist die Ähnlichkeitsmessung insofern *relational*, als entsprechende Funktionen $g \circ f : C \times C \to \mathbb{R}$, $f : C \times C \to X \times X$, $g : X \times X \to \mathbb{R}$, Paare von Textsegmenten auf Ähnlichkeitswerte abbilden. Der Einfachheit halber nehmen wir an, dass diese Werte im Intervall $I = [0,1]$ liegen und sich wie Zugehörigkeitswerte einer unscharfen Menge lesen lassen. Dieses Szenario lässt sich unmittelbar als Kategorisierungsaufgabe rekonstruieren und erlaubt daher den zuvor erarbeiteten evolutionären Zugriff. Dazu besteht die Möglichkeit einer Benennung von Teilintervallen des Intervalls I etwa mit Elementen der Menge $\mathscr{L}' = \{$*ähnlich*, *eher ähnlich*, *unentscheidbar*, *eher unähnlich*, *unähnlich*$\}$, so dass die Ähnlichkeitsmessung als Kategorisierungsfunktion $g' \circ f' : C \times C \to \mathscr{L}'$ notierbar wird. Man beachte, dass sich unüberwachte Verfahren der Textähnlichkeitsmessung dazu eignen, die durch $\mathscr{L}'$ benannte Partition mit Beispielen aufzufüllen – andernfalls ist mit intellektuell erzeugten Trainingsbeispielen zu operieren. Auf der Basis einer solchen Rekonstruktion kann nicht nur die Ähnlichkeitsmessung – wie bereits die Textkategorisierung – zum Gegenstand eines EA gemacht werden. Nach diesem Muster können vielmehr alle Klassen von Aufgaben, die binäre Relationen über Textsegmenten lernen (z. B. *Textual Entailment*), zum Gegenstand von VienNA gemacht werden.

4.3 Einbettung

An dieser Stelle sind mehrere Möglichkeiten denkbar, die Optimierung der Funktionen f und g in das Schema von Algorithmus 1 einzubetten. Wir beschränken uns auf die Variante, dass Algorithmen (R_1) und deren Parameter (R_2) gleichzeitig mit den vektoriellen Repräsentationen Input-bildender Texte (bzw. Textsegmente) (R_3) optimiert werden. Wegen der Größe von R_3 ist es wenig sinnvoll, die Suche in diesem Raum jener in R_1 (bzw. R_2) rekursiv unterzuordnen: Die evolutionäre Dynamik bliebe mutmaßlich auf R_3 beschränkt. Es besteht jedoch die Möglichkeit, in den Zeilen 6 und 7 aus Algorithmus 1 Bewegungen in R_3 wahrscheinlicher zu machen. Generationen von Lösungen (Zeile 1) entsprechen dann stets Parameterkonstellationen in allen drei Räumen ($R_1 \dots R_3$) und also Algorithmen und Repräsentationsmodellen, die in mehreren Teilpopulationen organisiert sein können, ob nun im Wettbewerb miteinander oder nicht (etwa dann, wenn sie dieselben oder verschiedene Klassifikationsaufgaben adressieren). Der entscheidende Punkt dabei ist, dass Parameterkonstellationen, die zu Klassifikationen von Textsegmenten und entsprechenden Annotationen führen, selbst zum Gegenstand der Merkmalsbildung im Sinne von Sektion 4.1 und 4.2 werden können.[18] Die Evaluationen der Zeilen 2 und 8 sind schließlich mittels einschlägiger Evaluationsmaße (z. B. F-Score) implementierbar, was initial bereitzustellende Trainings- und Testdaten voraussetzt.[19] Zeile 10 bietet den Einstiegspunkt für Wikiditoren, die nun alte oder neue Lösungen verwerfen bzw. auswählen können, was zu Revisionen von Evaluationen aus Zeile 8 führen kann. Die Selektionsfunktion aus Zeile 10 steht genauer in Zusammenhang mit drei Szenarien:

1. *Szenario 1:* Dies betrifft zunächst die bereits erläuterte Variante, bei der Wikiditoren die Selektion durchführen, z. B. zwecks Bedienung einer memetischen Funktion (siehe oben). Dieses Szenario entspricht einem Annotationsspiel, das auf zwei Ebenen überwacht ist: auf der Ebene der eingebauten Fitting-Funktionen (Zeile 2 und 8) und der durch Vorwissen oder Interessen geleiteten Selektion (Zeile 10), wobei durch Wikiditoren getätigte Bewertungen etwa in Form von Korrekturen das zugrundeliegende Trainings- und Testmaterial stetig erweitern bzw. verändern. Ein solches Anno-

---

18 Dadurch können beispielsweise Annotationen von Wortarten als Ergebnis des PoSTaggings anstelle der betroffenen Lexeme im Hinblick auf sequenzielle Muster betrachtet werden, die entsprechende Tokenfolgen erzeugen.

19 Von den hiermit zusammenhängenden Parameterräumen abstrahieren wir aus Platzgründen. Ferner sei darauf hingewiesen, dass dies zugleich der Ansatzpunkt für die anfangs der Sektion aufgezählten Fitness-Funktionen ist, welche weit über Maße wie F-Score hinauszugehen erfordern.

tationsspiel zielt auf folgende Leitfrage: *Inwieweit lösen die involvierten, fortwährend nachjustierten Verfahren und errechneten Repräsentationen die gestellten Aufgaben im Sinne der beteiligten Wikiditoren?*

2. *Szenario 2:* Demgegenüber besteht die Möglichkeit, die Selektion ebenfalls zu automatisieren, was zunächst auf Szenario 1 aus Tabelle 7.1 hinausliefe. Es besteht nun aber die Möglichkeit, die gestellten Annotationsaufgaben als Bots zu konzipieren, die sich variable Manifestationen in den Räumen R_1 ... R_3 selbstlernend suchen und einander dahingehend bewerten, inwieweit sie ihre Aufgabenbewältigung wechselseitig selbstorganisierend unterstützen. Ein solches Annotationsspiel zielt auf folgende Leitfrage: *Zu welchen Annotation führt „das freie Spiel" der Bots, die sich kooperationshalber mit welchen anderen Bots auf welche Weise vernetzen, so dass die Reihenfolge, in der diese agieren, zur variablen Modellgröße wird und nicht mehr als NLP-Pipeline vorgegeben zu werden braucht?*
3. *Szenario 3:* Ein drittes Szenario betrifft Mischformen letzterer beider Annotationsspiele, so dass Wikiditoren neben Bots als Selektoren in Wikidition agieren. Dieses Szenario zielt auf die Beantwortung folgender Leitfrage: *Ab welchem Zeitpunkt besteht die Möglichkeit des Übergangs zu Szenario 2, ohne dass Wikiditoren länger in den Selektionsprozess eingreifen müssen, um die evolutionäre Dynamik jenen Zustand approximieren zu lassen, auf den Szenario 1 zielt?* Szenario 3 adressiert sozusagen den Übergangsbereich von Szenario 1 zu 2.

Idealerweise liefe Szenario 3 darauf hinaus, die Annotation sprachlicher Einheiten, wie sie die Informatik und verwandte Disziplinen bislang unterstützen, soweit zu automatisieren, dass sie als Aufgabenstellung weitgehend erledigt wäre. Angesichts der Offenheit des menschlichen Interpretationsuniversums kann unmittelbar die Unmöglichkeit eines vollständigen Unterfangens dieser Art konstatiert werden. Dies wirft jedoch die Frage auf, wieweit man mit einer solchen Automatisierung gehen kann, inwieweit sich also die Arbeit der Computerlinguistik als wissenschaftliche Disziplin automatisieren lässt.

5 Schlussfolgerung

In diesem Beitrag haben wir vier Entwicklungslinien aufgezeigt, die einen erheblichen Anpassungsdruck auf die Weiterentwicklung von Infrastrukturen für die Digital Humanities erzeugen. Dies nahmen wir zum Anlass, mit TextImager, Wikidition und VienNA Beispiele für drei Klassen von Infrastrukturen zu betrachten, die sich unter anderem durch die Interaktionsmöglichkeiten

unterscheiden, die sie ihren Nutzern bieten. Mit VienNA haben wir an dritter Stelle eine neuartige Architektur für eine hochgradig interaktive Infrastruktur skizziert, die auf die vollständige Automatisierung der Annotation sprachlicher Einheiten zielt, so dass Interaktion letztlich die Kooperation selbstorganisierter, selbstlernender NLP-Bots meint. Zukünftige Arbeiten sind einer experimentellen Erprobung dieser Architektur gewidmet.

Literatur

Abokhodair, Norah, Daisy Yoo & David W. McDonald (2015): Dissecting a social botnet: Growth, content and influence in Twitter. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (CSCW '15), 839– 851. New York: ACM. doi: 10.1145/2675133.2675208.

Aggarwal, Charu C. & Jiawei Han (Hrsg.) (2014): *Frequent pattern mining*. Cham, Heidelberg, New York u. a.: Springer.

von Ahn, Luis (2008): Human computation. In *IEEE 24th International Conference on Data Engineering (ICDE 2008)*, 1–2. doi: 10.1109/ICDE.2008.4497403.

Bär, Daniel, Chris Biemann, Iryna Gurevych & Torsten Zesch (2012): UKP: Computing semantic textual similarity by combining multiple content similarity measures. In *Proceedings of SemEval '12*, 435–440. Stroudsburg: Association for Computational Linguistics.

Beißwenger, Michael & Angelika Storrer (2008): Corpora of computer-mediated communication. In Anke Lüdeling & Merja Kytö (Hrsg.), *Corpus Linguistics. An International Handbook of the Science of Language and Society*, Kap. 21, 292–309. Berlin/New York: de Gruyter.

Biemann, Chris, Gregory R. Crane, Christiane D. Fellbaum & Alexander Mehler (2014): Computational humanities – bridging the gap between computer science and digital humanities (Dagstuhl Seminar 14301). *Dagstuhl Reports* 4(7), 80–111. doi: 10.4230/DagRep.4.7.80.

Boshmaf, Yazan, Ildar Muslukhov, Konstantin Beznosov & Matei Ripeanu (2012): Key challenges in defending against malicious socialbots. In *Proceedings of the 5th USENIX Conference on Large-Scale Exploits and Emergent Threats* (LEET '12), 12–16. Berkeley: USENIX Association.

Brandes, Ulrik, Patrick Kenis, Jürgen Lerner & Denise van Raaij (2009): Network analysis of collaboration structure in Wikipedia. In *Proceedings of the 18th International Conference on World Wide Web (WWW '09)*, 731–740. New York, NY: ACM.

Eckart de Castilho, Richard & Iryna Gurevych (2014): A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In Nancy Ide & Jens Grivolla (Hrsg.), *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT) at COLING 2014*, 1–11. Dublin: Association for Computational Linguistics and Dublin City University.

Cheney-Lippold, John (2017): *We are data: Algorithms and the making of our digital selves*. New York: New York University Press.

Cresci, Stefano, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi & Maurizio Tesconi (2017): The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th International Conference on World Wide Web*

Companion WWW ’17 Companion, 963–972. Geneva: International World Wide Web Conferences Steering Committee. doi: 10.1145/3041021.3055135.

Cunningham, Hamish (2002): GATE, a general architecture for text engineering. *Computing and the Humanities* 36, 223–254.

Diesner, Jana (2014): ConText: Software for the integrated analysis of text data and network data (paper presented at the Social and Semantic Networks in Communication Research. Preconference at Conference of International Communication Association (ICA)), Seattle.

Ertek, Gurdal, Dilek Tapucu & Inanc Arin (2013): Text mining with RapidMiner. In Markus Hofmann & Ralf Klinkenberg (Hrsg.), *RapidMiner: Data Mining Use Cases and Business Analytics Applications*, 241–264. Boca Raton: CRC Press.

Everett, Daniel (2013): *Die größte Erfindung der Menschheit: Was mich meine Jahre am Amazonas über das Wesen der Sprache gelehrt haben*. DVA.

Ferrara, Emilio, Onur Varol, Clayton Davis, Filippo Menczer & Alessandro Flammini (2016): The rise of social bots. *Communications of the ACM* 59(7), 96–104. doi: 10.1145/2818717.

Ferron, Michela & Paolo Massa (2014): Beyond the encyclopedia: Collective memories in Wikipedia. *Memory Studies* 14(1), 22–45.

Freitas, Carlos, Fabrício Benevenuto, Adriano Veloso & Saptarshi Ghosh (2016): An empirical study of socialbot infiltration strategies in the Twitter social network. *Social Network Analysis and Mining* 6(1), 23.

Geiger, R. Stuart (2014): Bots, bespoke, code and the materiality of software platforms. *Information, Communication & Society* 17(3), 342–356.

Geiger, R. Stuart & David Ribes (2010): The work of sustaining order in Wikipedia: the banning of a vandal. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, 117–126. ACM.

Gilani, Zafar, Liang Wang, Jon Crowcroft, Mario Almeida & Reza Farahbakhsh (2016): Stweeler: A framework for Twitter bot analysis. In *Proceedings of the 25th International Conference Companion on World Wide Web* (WWW ’16 Companion), 37–38. Geneva: International World Wide Web Conferences Steering Committee. doi: 10.1145/2872518.2889360.

Gleim, Rüdiger, Alexander Mehler & Alexandra Ernst (2012): SOA implementation of the eHumanities Desktop. In *Proceedings of the Workshop on Service-oriented Architectures (SOAs) for the Humanities: Solutions and Impacts, Digital Humanities*, 1–6. Hamburg: Digital Humanities 2020. http://www.clarin-d.de/images/workshops/proceedingssoasforthehumanities.pdf (letzter Zugriff: 12. 12. 2017).

Guthrie, David, Ben Allison, Wei Liu, Louise Guthrie & Yorick Wilks (2006): A closer look at skip-gram modelling. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, 1222–1225.

Halfaker, Aaron & John Riedl (2012): Bots and cyborgs: Wikipedia’s immune system. *Computer* 45(3), 79–82.

Hearst, Marti A. (1999): Untangling text data mining. In *Proceedings of ACL ’99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland*, 3–10. Stroudsburg, PA: Association for Computational Linguistics.

Hearst, Marti A. (2017): eGaia, ein verteiltes, technisch-soziales, geistiges System. In John Brockmann (Hrsg.), *Was sollen wir von künstlicher Intelligenz halten? Die führenden Wissenschaftler unserer Zeit über intelligente Maschinen*, 338–339. Frankfurt a. M.: Fischer.

Hemati, Wahed, Tolga Uslu & Alexander Mehler (2016): TextImager: a distributed UIMAbased system for NLP. In *Proceedings of the COLING 2016 System Demonstrations*, 59–63. Osaka: Association for Computational Linguistics.

Hemati, Wahed, Tolga Uslu & Alexander Mehler (2017): TextImager as an interface to BeCalm. In Martin Krallinger & Alfonso Valencia (Hrsg.), *BioCreative V.5. Proceedings*, 47–53. Madrid: CNIO Centro Nacional de Investigaciones Oncológicas.

Hinrichs, Erhard, Marie Hinrichs & Thomas Zastrow (2010): Weblicht: Web-based LRT services for German. In *Proceedings of the ACL 2010 System Demonstrations* (ACLDemos '10), 25–29. Stroudsburg: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1858933.1858938.

Hollan, James, Edwin Hutchins & David Kirsh (2000): Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM Transaction on ComputerHuman Interaction* 7(2), 174–196.

Hoque, Enamul & Giuseppe Carenini (2015): ConVisIT: Interactive topic modeling for exploring asynchronous online conversations. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (IUI '15), 169–180. New York: ACM. doi: 10.1145/2678025.2701370.

Iosub, Daniela, David Laniado, Carlos Castillo, Mayo Fuster Morell & Andreas Kaltenbrunner (2014): Emotions under discussion: Gender, status and communication in online collaboration. *PloS ONE* 9(8), 1–23. doi: 10.1371/journal.pone.0104880.

Joachims, Thorsten (2002): *Learning to classify text using support vector machines*. Boston: Kluwer.

Kosko, Bart (2017): Denkmaschinen = alte Algorithmen auf schnelleren Computern. In John Brockmann (Hrsg.), *Was sollen wir von künstlicher Intelligenz halten? Die führenden Wissenschaftler unserer Zeit über intelligente Maschinen*, 338–339. Frankfurt a. M.: Fischer.

Kosorukoff, Alex (2001): Human based genetic algorithm. In *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, Band 5, 3464–3469. Tucson, AZ: Institute of Electrical and Electronics Engineers (IEEE). http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=972056 (letzter Zugriff: 12. 12. 2017).

Larsson, Anders Olof & Moe Hallvard (2015): Bots or journalists? News sharing on Twitter. *Communications* 40(3), 361–370.

Leuf, Bo & Ward Cunningham (2001): *The Wiki way. Quick collaboration on the web*. Boston: Addison Wesley.

Lobin, Henning (2014): *Engelbarts Traum: Wie der Computer uns Lesen und Schreiben abnimmt*. Frankfurt a. M.: Campus Verlag.

Lokot, Tetyana & Nicholas Diakopoulos (2016): News bots: Automating news and information dissemination on Twitter. *Digital Journalism* 4(6), 682–699.

Mehler, Alexander (2006): In search of a bridge between network analysis in computational linguistics and computational biology – a conceptual note. In Hamid R. Arabnia & Homayoun Valafar (Hrsg.), *Proceedings of the 2006 International Conference on Bioinformatics & Computational Biology (BIOCOMP '06), June 26, 2006, Las Vegas*, 496–500. Las Vegas: World Academy of Science.

Mehler, Alexander (2007): Compositionality in quantitative semantics. A theoretical perspective on text mining. In Alexander Mehler & Reinhard Köhler (Hrsg.), *Aspects of Automatic Text Analysis* (Studies in Fuzziness and Soft Computing), 139–167. Berlin, New York: Springer.

Mehler, Alexander (2010): Artifizielle Interaktivität. Eine semiotische Betrachtung. In Tilmann Sutter & Alexander Mehler (Hrsg.), *Medienwandel als Wandel von Interaktionsformen – von frühen Medienkulturen zum Web 2.0*, 107–134. Wiesbaden: Verlag für Sozialwissenschaften.

Mehler, Alexander, Rüdiger Gleim, Tim vor der Brück, Wahed Hemati & Tolga Uslu (2015): Wikidition: Automatic lexiconization and linkification of text corpora. Submitted.

Mehler, Alexander, Rüdiger Gleim, Wahed Hemati & Tolga Uslu (2018): Skalenfreie online soziale Lexika am Beispiel von Wiktionary. In Stefan Engelberg, Henning Lobin, Kathrin Steyer & Sascha Wolfer (Hrsg.), *Jahrbuch des Instituts für Deutsche Sprache 2017*, Berlin: de Gruyter.

Mehler, Alexander, Serge Sharoff & Marina Santini (Hrsg.) (2010): *Genres on the web: Computational models and empirical studies*. Dordrecht: Springer.

Mehler, Alexander, Benno Wagner & Rüdiger Gleim (2016): Wikidition: Towards a multilayer network model of intertextuality. In *Proceedings of Digital Humanities 2016*, Kraków: Alliance of Digital Humanities Organizations.

Michelucci, Pietro (2013): Synthesis and taxonomy of human computation. In Pietro Michelucci (Hrsg.), *Handbook of Human Computation*, 83–86. New York: Springer.

Milo, R., S. Shen-Orr, S. Itzkovitz, N. Kashtan & D. Chklovskiiand U. Alon (2002): Network motifs: Simple building blocks of complex networks. *Science* 298(5594), 824–827.

Nickerson, Jeffrey V. (2013): Human-based evolutionary computing. In Pietro Michelucci (Hrsg.), *Handbook of Human Computation*, 641–648. New York: Springer.

Patel, Kayur, Naomi Bancroft, Steven M. Drucker, James Fogarty, Andrew J. Ko & James Landay (2010): Gestalt: Integrated support for implementation and analysis in machine learning. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology* (UIST ’10), 37–46. New York: ACM. doi: 10.1145/1866029.1866038.

Peng, Roger D. (2011): Reproducible research in computational science. *Science* 334(6060), 1226–1227.

Rockwell, Geoffrey & Stéfan Sinclair (2016): *Hermeneutica: Computer-assisted interpretation in the humanities*. Cambridge, MA: MIT Press.

Stegbauer, Christian (2009): *Wikipedia: Das Rätsel der Kooperation*. Wiesbaden: Verlag für Sozialwissenschaften.

Steiner, Thomas (2014): Telling breaking news stories from Wikipedia with social multimedia: A case study of the 2014 winter olympics https://arxiv.org/abs/1403.4289 (letzter Zugriff: 12. 12. 2017).

Subrahmanian, V. S., Amos Azaria, Skylar Durst, Vadim Kagan, Aram Galstyan, Kristina Lerman, Linhong Zhu, Emilio Ferrara, Alessandro Flammini & Filippo Menczer (2016): The DARPA Twitter bot challenge. *Computer* 49(6), 38–46.

Tapscott, Don & Anthony D. Williams (2008): *Wikinomics: How mass collaboration changes everything*. New Jork: Portfolio.

Uslu, Tolga, Wahed Hemati, Alexander Mehler & Daniel Baumartz (2017): TextImager as a generic interface to R. In *Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, 17–20. Valencia: Association for Computational Linguistics.

Wilks, Yorick, Roberta Catizone, Simon Worgan, Alexiei Dingli, Roger Moore, Debora Field & Weiwei Cheng (2010): A prototype for a conversational companion for reminiscing about images. *Computer Speech and Language* 25, 140–157.

Zaki, Mohammed J. (2005): Efficiently mining frequent trees in a forest: Algorithms and applications. *IEEE Transactions on Knowledge and Data Engineering* 17(8), 1021–1035. doi: 10.1109/TKDE.2005.125.

Hans-Jürgen Bucher und Philipp Niemann

8 Infrastrukturen zur Erforschung medienspezifischer Sprachverwendung

Am Beispiel der Rezeptionsforschung

Abstract: Infrastrukturen in der Medienforschung haben sich lange Zeit auf die Dokumentation und Speicherung von Print-, Audio- und Video-Materialien beschränkt, die in Kooperation mit technischen Abteilungen der Universitäten oder Forschungseinrichtungen bewerkstelligt wurden. Durch die experimentelle Medienforschung und die Beforschung großer Internet-Korpora ergeben sich neuartige infrastrukturelle Anforderungen, auf die auch die entsprechenden Förderprogramme nicht optimal abgestimmt sind. Der Beitrag stellt die Chancen und Probleme gerätebasierter Medienforschung für zwei Forschungsbereiche dar: die experimentelle Datenerhebung in der Rezeptionsforschung und die digitalisierte Ablage von Forschungsdaten.

Keywords: Experimentelle Datenerhebung, Forschungsförderung, Forschungswerkzeuge, Medienwissenschaft, Rezeptionsforschung

1 Die Ausgangslage: zum Bedarf an Infrastrukturen in der Medienforschung

In den Geistes- und Kulturwissenschaften nahm die sprachwissenschaftliche Medienforschung hinsichtlich spezifischer infrastruktureller Anforderungen für lange Zeit eine Sonderstellung ein: Zeitungsforschung war auf Vervielfältigungs- und Archivierungstechnologien – z. B. Mikrofilme – angewiesen, Fernsehforschung, Hörfunkforschung und später die Online-Forschung erforderten elektronische Aufzeichnungs- und Speichertechnologien. Diese infrastrukturellen Anforderungen wurden in der Regel durch eine Kooperation zwischen

Hans-Jürgen Bucher, Medienwissenschaft, Universität Trier, Universitätsring 15, D-54286 Trier, E-Mail: bucher@uni-trier.de
Philipp Niemann, Abteilung Wissenschaftskommunikation, Institut für Germanistik, Karlsruher Institut für Technologie (KIT), Kaiserstr. 12, D-76131 Karlsruhe, E-Mail: philipp.niemann@kit.edu

https://doi.org/10.1515/9783110538663-009

Wissenschaft und technischen Abteilungen der Universitäten oder der Forschungseinrichtungen sichergestellt. Was allerdings die Analyse des gespeicherten Materials betrifft, so ist die sprachwissenschaftliche Medienforschung, wie die linguistische Forschung zur Sprachverwendung insgesamt, lange Zeit weitgehend ohne infrastrukturelle Ausstattung ausgekommen. Empirische Forschung bedeutete in der Regel das Sammeln aussagekräftiger Beispiele, das Anlegen von kleinen Korpora und die händische Analyse exemplarischer Fälle, die dazu beitragen sollten, den Zusammenhang zwischen Medienstrukturen einerseits und Sprach- und Kommunikationsstrukturen andererseits zu erhellen. Es dominierten qualitative, interpretierende Methoden, die von Einzelforschern oder kleineren Forschergruppen leicht zu organisieren waren. Aufgrund der Spezifik der verschiedenen Fallstudien hat die Frage nach möglichen Sekundäranalysen der Daten und Materialien ebenso keine relevante Rolle gespielt wie die Vergleichbarkeit der Befunde. Wie die in den verschiedenen Sammelwerken (Leonhard et al 2002) und Übersichtsdarstellungen (Burger & Luginbühl 2014; Bucher 2014) präsentierten Forschungsbefunde zeigen, konnte auf diese Art eine ganze Reihe wichtiger Einsichten in die Medienspezifik von Sprache und Kommunikation gewonnen werden. Mit Ausnahme einiger Studien zur Online-Kommunikation (z. B. Bucher 2009; Thimm, Einspänner & Dang-Anh 2012; Dang-Anh & Rüdiger 2015) sind begrenzte, exemplarische Fallstudien auch heute noch die Standardforschungsstrategie. Die mit einem solchen Forschungsansatz gewonnenen Einsichten sind allerdings meistens für die Sprachwissenschaft deutlich relevanter als für die Medien- und Kommunikationswissenschaft. Deren Erkenntnisinteresse ist weniger auf exemplarische Einzelbefunde gerichtet, als vielmehr auf strukturelle Befunde über größere Datenmengen, auf Rezeption und Wirkungen medialer Kommunikation oder Erklärungen für die Entstehung massenmedialer Kommunikationsbeiträge.

Eine Ausnahme stellen in dieser Hinsicht die wenigen Ansätze einer computerlinguistischen Medienforschung dar: Sowohl für die Datenvorhaltung in Form von digitalen Korpora als auch für die Datenanalyse kommen informationstechnologische Infrastrukturen zum Einsatz (Beißwenger 2007). Allerdings haben die Analysepotentiale der Computerlinguistik die medienwissenschaftliche Forschung bislang erst ansatzweise befruchtet und damit in der Medienforschung auch keine Notwendigkeit zur Etablierung entsprechender Infrastrukturen geschaffen. Einer der wesentlichen Gründe dafür liegt in der Unverträglichkeit zweier Forschungslogiken: Während die Computerlinguistik für die Bearbeitung ihrer Fragestellungen auf annotierte Korpora und Metadaten angewiesen ist, bestehen die Forschungsgegenstände in der Medien- und Kommunikationswissenschaft gerade aus unstrukturierten Datensammlungen, die auch noch verschiedene nicht-sprachliche bedeutungsrelevante

Modi enthalten wie Abbildungen, Farben, Design, Musik oder Typografie. Eine Annotierung solcher unstrukturierter Datenbestände würde den Zeithorizont medien- und kommunikationswissenschaftlicher Forschung sprengen, so dass eine Integration computerlinguistischer Verfahren in der Medienforschung bislang eher die Ausnahme geblieben ist. Man kann vor dem Hintergrund der skizzierten historischen Entwicklung die These vertreten, dass ein Ausbau der Forschungsinfrastruktur eine Voraussetzung dafür ist, dass die sprach- und diskursanalytische Medienforschung Anschluss an die sozialwissenschaftliche Medienforschung findet, denn um zu strukturellen Befunden über größere Datenmengen zu gelangen, bedarf es nicht nur einer Erweiterung des Forschungs-Know-hows, sondern auch einer IT-Infrastruktur, die computerbasierte Analysen ermöglicht.

Eine frühe Ausnahme hinsichtlich einer Verbindung medienwissenschaftlicher, sprachwissenschaftlicher und computergestützter Verfahren ist das an der Universität Tübingen Anfang der 1990er Jahre durchgeführte DFG-Projekt *Die Sprache der ersten deutschen Wochenzeitungen im 17. Jahrhundert*. In diesem Projekt wurde das zu analysierende Textkorpus händisch aus den Originaltexten erfasst, in eine relationale Datenbank überführt und lexikalisch, syntaktisch und funktional annotiert. Dadurch war es möglich, für die fünf untersuchten Zeitungen Konkordanzen zu erstellen und die Zeitungstexte computergestützt in Bezug auf Wortschatz, Syntax, Textstrukturen und Darstellungsformen zu analysieren. Auf diese Weise ließen sich qualitative Fallstudien und quantitative Auswertungen in einem für die damalige Zeit außergewöhnlichen Forschungsdesign integrieren (Fritz & Straßer 1996; Schröder 1995). Die Infrastruktur des Projektes hat sich dabei auf Einzelrechner mit einer Datenbank und einem Textverarbeitungsprogramm beschränkt.

2 Digitalisierung als neue Herausforderung für wissenschaftliche Infrastrukturen in der Medienforschung

Mit der Digitalisierung haben sich die Anforderungen an eine IT-Infrastruktur auch in den Geistes- und Kulturwissenschaften grundlegend geändert. So heißt es in der *Stellungnahme der Kommission für IT-Infrastruktur für 2016–2020* der Deutschen Forschungsgemeinschaft in Bezug auf die Geisteswissenschaften:

> Datengetriebene Wissenschaft führt zu neuen Erkenntniswegen. Neben den Möglichkeiten zur Zugänglichmachung von Artefakten, z. B. in Form der Erfassung alter Schriften,

> Bilder etc., bieten die Analyse- und Auswertungsmethoden zur schnellen Verarbeitung riesiger Datenmengen – Stichwort „Big Data“ – neue Mittel zum Erkenntnisgewinn. (DFG 2016: 5)

Auch der Wissenschaftsrat sieht eine Transformation der Forschungsinfrastruktur

> von tradierenden und Fachinformationen bevorratenden Hilfseinrichtungen zu Inkubatoren für neue und innovative wissenschaftliche Fragestellungen aufgrund von Forschungsdaten, die durch diese Infrastrukturen selbst erst erzeugt werden (Wissenschaftsrat 2011).

Der *Strategy Report on Research Infrastructures* von 2016 des *European Strategy Forums on Research Infrastructures* (ESFRI) der EU-Kommission bestätig diese Entwicklung auch im internationalen Vergleich: „The increased availability of digital resources and the development of advanced digital methods for research in the Humanities have heralded remarkable changes in the scale and scope of research in these disciplines.“ (ESFRI 2016: 170)

Für die Medien- und Kommunikationswissenschaft bedeutet diese Entwicklung in erster Linie die Verfügbarkeit von größeren, digitalen Datenbeständen, die zuvor nur in analoger Form als Texte (Printprodukte, Transkripte) oder audio-visuelle Aufzeichnungen vorlagen. Allerdings erfolgt die Analyse digitaler Gegenstände bis heute nahezu ausschließlich durch händische Kodierung. Digitale Auswertungsmethoden – beispielsweise durch die Statistik-Software SPSS – kommen erst nach der händischen Kodierung zum Einsatz. Eine Ersetzung oder zumindest eine Ergänzung der händischen Kodierung könnte die Kooperation zwischen computerlinguistischen und sozialwissenschaftlichen Ansätzen leisten. Die dafür erforderliche Entwicklung automatisierter, computerbasierter Analysemethoden würde allerdings auch neuartige Anforderungen an eine Forschungsinfrastruktur stellen. Im Hinblick auf eine solche integrative Entwicklung diagnostiziert das Strategie-Forum der Europäischen Kommission bereits eine Auflösung der Grenzen zwischen verschiedenen Wissenschaftskulturen:

> On the one hand, wide corpora of digitised texts allow Humanities to use quantitative methods that were previously confined to social sciences. On the other hand, a “linguistic turn” of the social sciences, makes room for new types of discourse and conversation analysis. Media Studies, which connect the Social Sciences and Humanities, are an eloquent example of that evolution. In particular, the scientific study of the web, which has become an integrated part of society, culture, business, and politics, is a burgeoning field of research activity, with enormous potential for the contribution of SSH to societal challenges relating to communication or security. (ESFRI 2016: 172)

Potentiale und Anforderungen an Forschungsinfrastrukturen werden deutlicher erkennbar, wenn man sie in Funktionsbereiche unterteilt, wie sie auch

von Seiten der Forschungsförderung untergliedert werden (vgl. Wissenschaftsrat 2011, ESFRI 2016).

1. Großgeräte. z. B. Teilchenbeschleuniger, Teleskope, große Laborgeräte;
2. Forschungsinformationsinfrastrukturen, wie Sammlungen, Archive, Digitale Datenbanken;
3. informationstechnische Infrastrukturen oder e-Infrastrukturen, wie Groß- und Hochleistungsrechner, Hochleistungskommunikations- und Rechnerverbünde;
4. soziale Forschungsinfrastruktur, wie Begegnungsräume des diskursiven Austauschs von aktuellen und der Entwicklung von neuen Forschungsfragen.

In seiner Bestandsaufnahme, die sich auf eine Umfrage unter 99 Fachgesellschaften der Sozial- und Geisteswissenschaften stützt, kommt der Wissenschaftsrat zu dem Ergebnis, dass es sich bei über 90 % der infrastrukturellen Einrichtungen um digitale Ressourcen wie Datenbanken oder Textkorpora handelt, und weniger als 10 % auf technische und räumliche Ressourcen – also Großgeräte – entfallen.

Dass die Medienwissenschaft ein Bindeglied zwischen den Humanities und den Sozialwissenschaften sein kann, drückt sich auch darin aus, dass für die Medienforschung sowohl digitale Ressourcen als auch Großgeräte neue Forschungsfragen eröffnen können. Digitale Ressourcen sind z. B. erforderlich für die Speicherung und Erschließung großer Mediendaten aus Printmedien oder audiovisuellen Medien. Insbesondere die Online-Forschung setzt voraus, dass große, bereits digital vorliegende Internet-Datenbestände für eine langfristige Erschließung vorgehalten werden können. Eine Datenbank im Fach Medienwissenschaft der Universität Trier von Tweets zu rund 200 Stichwörtern, gesammelt über einen Zeitraum von zwei Jahren, umfasst beispielsweise rund 60 Millionen Tweets mit einem Speichervolumen von mehreren 100 GB. Die Erschließung und Auswertung dieser Daten erfordert eine digitale Infrastruktur mit einer entsprechenden Auswertungssoftware und Speichermedien für die Vorhaltung der Datenbestände, was von einem einzelnen Fach nur in Kooperation mit zentralen Infrastruktureinrichtungen, wie einem Rechenzentrum, zu leisten ist und entsprechende Folgekosten verursacht. Ein Ausbau der Infrastruktur in den Geistes- und Sozialwissenschaften, der nachhaltig sein soll, erfordert dementsprechend über die technische Anschaffung hinaus zusätzliche Personalmittel für Betrieb und Betreuung sowie etatisierte Finanzierungsmittel für Folgekosten wie Server oder Speicherplatz.

Im Folgenden werden Herausforderungen für eine Forschungsinfrastruktur in der Medienforschung für zwei Forschungsbereiche vorgestellt und diskutiert: für den Bereich der großgeräte-basierten Datenerhebung und den Bereich der Datenablage und -vorhaltung.

3 Infrastrukturen für die Datenerhebung

3.1 Großgeräte für die empirische Medienforschung: Probleme der Finanzierung

Großgeräte für experimentelle Laborforschung sind in der Medien- und Kommunikationswissenschaft ebenso wie in den Sozialwissenschaften insgesamt bislang eher die Ausnahme geblieben. Das belegt auch die Verteilung der entsprechenden Fördermittel. Trotz der auf nationaler und internationaler Ebene erhobenen Forderung nach einem Paradigmenwechsel in den Forschungsstrategien von Sozial- und Geisteswissenschaften ist die finanzielle Förderung in diesen Disziplinen immer noch weit hinter den Natur-, Ingenieurs- und Lebenswissenschaften zurückgeblieben. Das zeigt sich, wenn man die Bewilligungen in den beiden wichtigsten nationalen Förderprogrammen für Großgeräte nach Fachdisziplinen vergleicht: das Förderprogramm *Forschungsgroßgeräte* der DFG nach Art. 91b GG und das Programm *Großgeräte der Länder*.

Von den über 1.500 bewilligten Forschungsgroßgeräten durch die DFG stammen gerade acht Anträge aus den Sozial- und Geisteswissenschaften, bei den Großgeräten der Länder waren es 89 von 1389. Von den über 800 Millionen Euro an Förderung für Großgeräte entfallen 2,5 Millionen auf die Sozial- und Geisteswissenschaften und beispielsweise über 300 Millionen Euro auf die Naturwissenschaften (DFG: Fünf Jahre neue Großgeräteprogramm 2007–2011: 21, Tab. 5 und 6). Natürlich ist diese Diskrepanz mit den unterschiedlichen Anforderungen an eine Forschungsinfrastruktur in den verschiedenen Wissenschaftskulturen zu erklären. So werden in den Geistes- und Sozialwissenschaften keine Massenspektrometer, Laserscanning-Mikroskope oder Rasterelektronenmikroskope mit einem Anschaffungswert von mehreren hunderttausend Euro benötigt. Die von der DFG festgelegte Bagatellgrenze für Großgeräte – ganz unabhängig von verschiedenen Forschungstraditionen – von 200.000 Euro begünstigt jedoch die Fachdisziplinen, die solche aufwendigen Geräte benötigen, gegenüber den Geistes- und Sozialwissenschaften, deren infrastrukturelle Anforderungen mit deutlich geringeren Summen befriedigt werden könnten. In seinen *Empfehlungen zur Forschungsinfrastrukturen in den Geistes- und Sozialwissenschaften* fordert dementsprechend der Wissenschaftsrat, deren Besonderheiten gegenüber den Naturwissenschaften bei der Förderung zu berücksichtigen und „von Bagatellgrenzen bei Investitionskosten für die Geistes- und Sozialwissenschaften abzusehen“ (Wissenschaftsrat 2011: 11).

Neben der Bagatellgrenze liegt ein weiteres Problem für die Großgeräteausstattung sozialwissenschaftlicher Disziplinen darin, dass ihre Anträge bei der DFG nach denselben Kriterien beurteilt werden, wie die aus den klassischen

Großgeräte-Disziplinen. Das betrifft beispielsweise die Bewertung der Publikationsleistungen von Antragstellern: Während in den Naturwissenschaften oder den Lebenswissenschaften die Zeitschriftenpublikation den Normalfall darstellt, sind in den Sozial- und Geisteswissenschaften auch Buch- und Sammelband-Veröffentlichungen etablierte Formen der internen Wissenschaftskommunikation. Solange letztere in der Publikationsbewertung niedriger eingeschätzt werden, bleiben die entsprechenden Fachtraditionen benachteiligt. Während im Zusammenhang von e-Learning-Initiativen Großgeräte bei der DFG erfolgreich beantragt werden konnten, ist das für sozialwissenschaftliche Laborausstattungen immer noch schwierig: da experimentelle Laborforschung in der Medien- und Kommunikationswissenschaft nicht zu den zentralen Forschungsfeldern gehört, bedarf es für die Antragsteller besonderer Anstrengungen, ihre Kompetenzen in diesem Bereich nachzuweisen. Eine weitere Schwierigkeit der Antragstellung liegt in der Großgeräte-Definition der DFG: „Ein Großgerät ist die Summe der Geräteteile einschließlich Zubehör, die für einen vorgesehenen Betriebszustand eine Funktionseinheit bildet" (http://www.dfg.de/foerderung/programme/infrastruktur/wgi/). Der Nachweis, dass ein Rasterelektronenmikroskop in der Geophysik im „vorgesehenen Betriebszustand eine Funktionseinheit bildet", ist gegenüber einer komplexen und mehrteiligen Laborausstattung sicher einfacher zu führen. Eine solche Laborausstattung zeichnet sich beispielsweise dadurch aus, dass sie aus mehreren Komponenten zusammengesetzt sein kann, die in einer zeitlichen und variablen Abfolge zum Einsatz kommen. So kann bereits die Datenerhebung aus verschiedenen Schritten bestehen und verschiedene Einzelgeräte erfordern: Aufmerksamkeitsmessungen mittels Eye-Tracker, Befragung am Bildschirm, Leitfadeninterviews mit audiovisueller Dokumentation. Entsprechend der experimentellen Forschungslogik besteht eine Laborausstattung in der Regel aus einem Datenerhebungssystem, einem Auswertungs- und Aufbereitungssystem, die nicht alle gleichzeitig eine Funktionseinheit bilden, sondern sequentiell eingesetzt werden: erst die Datenerhebung, dann die Aufbereitung und schließlich die Datenauswertung. Die geschilderten Unverträglichkeiten zwischen Fördervorgaben und Forschungspraxis sind charakteristisch für eine ganze Reihe sozial- und kulturwissenschaftlicher Großgeräteanträge, weshalb diese gegenüber naturwissenschaftlichen Anträgen ein strukturell bedingtes hohes Ablehnungsrisiko aufweisen.

3.2 Potentiale der Medienforschung mit Großgeräten: Beispiel Rezeptionslabor

Fragen des Verstehens und der Verständlichkeit haben in der Sprachwissenschaft nicht nur unter theoretischen Gesichtspunkten eine wichtige Rolle ge-

spielt, sondern bereits sehr früh auch im Hinblick auf eine Anwendungsorientierung sprachwissenschaftlicher Forschung (Muckenhaupt 1980; Feuerstein & Heringer 1987; Spillner 1995). Experimentelle Forschung dazu hat aber in der Linguistik ebenso wenig stattgefunden wie in der an Medienwirkungen interessierten Massenkommunikationsforschung. Während erstere Verstehen und Verständlichkeit kommunikationsanalytisch und materialbezogen untersucht, rekonstruiert die Massenkommunikationsforschung Medienwirkungen mittels einer Integration von Inhaltsanalyse und Umfragedaten. Ein Rezeptionslabor schafft die Voraussetzung, die Interaktion zwischen einem medialen Stimulus und einem Rezipienten experimentell und direkt zu erforschen und damit die Grundlage sowohl für Medienwirkungsbefunde als auch für Einsichten in Verstehensprozesse oder Kriterien der Verständlichkeit zu schaffen. Am Beispiel des Rezeptionslabors der Trierer Medienwissenschaft soll exemplarisch gezeigt werden, welche Potentiale eine entsprechende Großgeräte-Infrastruktur sowohl für die Grundlagenforschung als auch für die angewandte Forschung schaffen kann.

Das Rezeptionslabor der Trierer Medienwissenschaft wurde in der Erstausstattung 2003 auf Grundlage eines Großgeräteantrags nach Art. 91b GG an die DFG angeschafft und 2010 mit Mitteln des Landes Rheinland-Pfalz modernisiert. Die Antragstellung erfolgte interdisziplinär unter Beteiligung der Medienwissenschaft, der Soziologie, der Politikwissenschaft, der Kunstgeschichte, der Wirtschaftswissenschaft und der Ethnologie. Die Ausstattung umfasst zwei Blickaufzeichnungskomponenten: eine Remote-Ausstattung für bildschirmbasierte Blickaufzeichnungen und eine sogenannte „Brille" (*glasses*) für szenische Blickaufzeichnungen. Letztere wurde für Blickaufzeichnungen beim Zeitungs- und Zeitschriftenlesen (Bucher, Schumacher & Duckwitz 2007), für die Rezeption wissenschaftlicher Vorträge mit Präsentationen (Bucher, Niemann & Krieg 2010) sowie für Rezeptionsuntersuchungen von Museumsbesuchen genutzt. Die Remote-Ausstattung wurde eingesetzt für Rezeptionsstudien zu Internetseiten, zu Filmen und Videos (Bucher 2011; Bucher, Schumacher & Duckwitz 2012) zu Wissens-Comics (Bucher & Boy 2018) oder zu digitalisierten Zeitungs- und Zeitschriftenausgaben (Bucher & Schumacher 2006, Gehl 2013). Einige der Studien waren als Grundlagenforschung angelegt, um beispielsweise für multimodale Stimuli die Integration verschiedener Kommunikationsmodi wie Sound, Musik, Sprache, Text oder Design im Rezeptionsprozess zu erforschen (Bucher & Niemann 2012; Bucher & Schumacher 2006; Bucher 2012, 2017). Andere Studien waren stärker anwendungsorientiert ausgerichtet, beispielsweise Usability-Studien zu verschiedenen Online-Angeboten (Bucher 2002) und interaktiven Darstellungsformen (Schumacher 2009), oder Studien zur Ermittlung des Wissenstransfers durch verschiedene Zeitschriften-

typen, wissenschaftliche Vorträge (Bucher & Niemann 2015) oder Informationsgrafiken (Gehl 2013). Die grundlagentheoretischen Erkenntnisse sowie einige der anwendungsorientierten Forschungsbefunde sind in einem Sammelband dokumentiert, der auch als Grundlage für weitere Forschungsarbeiten dient und als Standardwerk für eine interaktionsanalytische Rezeptionsforschung gelten kann (Bucher & Schumacher 2012). Neben der Forschung wurde das Rezeptionslabor auch in der Lehre eingesetzt, um die Studierenden der Medienwissenschaft einerseits mit den experimentellen Methoden vertraut zu machen, damit sie diese in eigenständigen Forschungsprojekten anwenden können, und um andererseits berufsbezogene Qualifikationen zu vermitteln, die auf Berufsfelder in der kommerziellen Online- oder Medienforschung vorbereiten.

Geräteinfrastruktur erfordert allerdings auch eine entsprechende Personalstruktur: Neben der technischen Betreuung der Anlage muss auch gewährleistet sein, dass das entsprechende Know-how für die Entwicklung von zielführenden Forschungsdesigns, die Datenerhebung und Auswertung, die Behebung von Störungen und Softwareabstürzen, die Expertenkommunikation mit dem Geräteanbieter, die Schulung von Studierenden und Forschern sowie für die Marktbeobachtung hinsichtlich Geräteinnovationen vorgehalten werden kann. Da sozial- und kulturwissenschaftliche Einrichtungen oder Abteilungen in der Regel gerade nicht mit Personal für derartige Aufgaben ausgestattet sind, bleibt es schwierig, außerhalb von geförderten Projekten mit entsprechenden Personalmitteln die Funktionsfähigkeit eines Labors konstant zu erhalten. Insofern generiert ein Labor den Zwang zur kontinuierlichen Einwerbung von Projektmitteln oder aber zur Zusatzbelastung des festangestellten Personals. Die Kontinuität der empirischen Rezeptionsforschung in Trier sowie die Aufrechterhaltung der Funktionsfähigkeit des Labors konnte nur durch eine Kombination dieser beiden Strategien über nunmehr fast 15 Jahre sichergestellt werden. Eine personelle Unterversorgung oder eine Verlagerung der genannten Aufgaben auf studentische Hilfskräfte führt in der Regel früher oder später zu einer Auflösung des Labors.

Eine vorbildliche institutionalisierte Lösung dieser Probleme bietet das Humanities Lab an der Universität Lund, das in seiner Art einzigartig in Europa ist. Das Labor ist eine eigenständig zentrale Abteilung der geisteswissenschaftlichen und der Theologischen Fakultät mit eigener Leitung und Budgetierung und mit 25 Beschäftigten in Forschung und Verwaltung (Stand: Frühjahr 2017). Die Funktion des Labors ist es „to promote the diversification of research in the Humanities and Theology“, es soll als offene Einrichtung Forscher aus verschiedenen geisteswissenschaftlichen Disziplinen dabei unterstützen, traditionelle und innovative Forschungsmethoden zu kombinieren, und es soll als Interface fungieren zwischen „academia and external actors in industry and

education“ (http://www.humlab.lu.se/en/about/policy-documents/ letzter Zugriff: 24. 11. 2017). In der Selbstdarstellung heißt es:

> The Humanities Lab is an interdisciplinary department for research technology and training at the Joint Faculties of Humanities and Theology. We host technology, methodological know-how, archiving expertise, and a wide range of research projects. Lab activities are centered around the humanities with research targeting issues of communication, culture, cognition and learning, but many projects are interdisciplinary and conducted in collaboration with the social sciences, medicine, the natural sciences, engineering, and e-Science (http://www.humlab.lu.se/en/about/, letzter Zugriff: 6. 11. 2017).

Neben einer Blickaufzeichnungs-Ausstattung mit allen derzeit relevanten Technologien umfasst das Labor eine Artikulographie zur Erforschung der Lautproduktion, Systeme für Hautwiderstands-, Herzschlag- und Atemmessungen, ein Elektroenzephalogramm zur neurologischen Messung von Hirnaktivitäten, ein Motion-Capture-System zur Untersuchung von Bewegungen und Gesten sowie ein Virtual-Reality-Studio. Mit dieser Ausstattungsbreite sind nahezu alle rezeptiven Prozesse menschlicher Kommunikation sowohl in Face-to-Face- als auch in Face-to-Interface-Konstellationen erfassbar. Durch die strukturelle Verankerung des Labors als zentrale Einrichtung sowie durch regelmäßige Kursangebote ist sichergestellt, dass seine Funktionalitäten auch von Wissenschaftlern genutzt werden können, die bislang keine experimentelle Forschung betrieben haben. Der wissenschaftliche Output des Labors sowie sein internationaler Ruf als Ausbildungs- und Tagungsstätte belegen die Produktivität des Konzeptes. Aufgrund seiner profilierten Stellung ist das Lunder Humanities Lab inzwischen auch zu einem Kooperationspartner der entsprechenden Gerätehersteller geworden und kann diese dabei unterstützen, wissenschaftliche Anforderungen frühzeitig in die Geräte-Entwicklung einzubeziehen.

Am Beispiel des Lunder Humanities Lab wird deutlich, dass Infrastrukturmaßnahmen nicht auf die technische Ausstattung beschränkt werden können, sondern personelle Ressourcen und strukturelle Reorganisationen umfassen müssen. Solange die Geistes- und Sozialwissenschaften auch in dieser Hinsicht den Natur- und Lebenswissenschaften hinterherhinken, wird auch eine technische Aufrüstung durch Großgeräte nicht zu den gewünschten Forschungserfolgen führen können.

4 Datenablage: Qualitätssicherung durch Standardisierung

Wer in der Medien- und Kommunikationswissenschaft heute qualitative Rezeptionsforschung betreibt, der archiviert und dokumentiert seine Daten in aller

Regel nach eigenem Ermessen und nach eigenen Prinzipien. Standardisierungen über Projekt- oder Lehrstuhlgrenzen hinaus sind die Ausnahme. Was das GESIS – Leibniz-Institut für Sozialwissenschaften für die quantitative Sozialforschung leistet, fehlt in der qualitativen Rezeptionsforschung bislang: die langfristige Sicherung von Forschungsdaten, ihre einheitliche Dokumentation nach festen Standards und die Bereitstellung der Daten für Forschung und Öffentlichkeit (vgl. GESIS – Leibniz-Institut für Sozialwissenschaften e. V. o. J.). Dass darin ein Manko gesehen werden muss, liegt auf der Hand: Der zunehmende Druck des Relevanznachweises gegenüber Entscheidungsträgern aus Wissenschaft und Politik und nicht zuletzt auch der breiten Öffentlichkeit trifft die qualitative Forschung nicht nur in gleicher Weise wie ihr quantitatives Pendant. Die Problematik ist hier sogar insofern virulenter, als auch innerhalb der Scientific Community jede Forschung, die das Verstehen von Phänomenen im primären Fokus hat und nicht mit Begriffen wie „signifikant“ oder „repräsentativ“ aufwarten kann, ihre Daseinsberechtigung stets aufs Neue belegen muss.

Nichts läge also näher, als durch Standardisierungsbemühungen auf der Ebene der Dokumentation, Ablage und Zugänglichkeit von Forschungsdaten der qualitativen Rezeptionsforschung dem zentralen Grundprinzip wissenschaftlichen Arbeitens, der intersubjektiven Nachvollziehbarkeit von Forschungsprozessen und -ergebnissen (Stichwort Replizierbarkeit, vgl. auch Huschka & Oellers 2013: 10) zu entsprechen und damit neben der Sicherung von Forschungsqualität auch noch der gesellschaftlichen Forderung nach Transparenz und Zugänglichkeit von Daten nachzukommen. Dass in diesem Bereich dennoch momentan in Deutschland keine Bestrebungen sichtbar sind, hängt mit den etablierten Arbeitsweisen und den Untersuchungsgegenständen im Bereich der qualitativen Rezeptionsforschung zusammen: Die Forschungseinheiten sind in der Regel klein. Anders als in den Naturwissenschaften sind Arbeitsgruppen mit mehreren Wissenschaftlern und einer Reihe von Doktoranden und Studierenden die absolute Ausnahme. Die Arbeit findet auf der Ebene von Einzellehrstühlen mit wenigen Mitarbeitern statt, größere Projektverbünde, Forschergruppen oder Netzwerke, die sich mit vergleichbaren Theorien und Methoden befassen, finden sich in der qualitativen Rezeptionsforschung nicht. Das bringt es mit sich, dass auch die Gegenstände der Forschung häufig nur von einem Lehrstuhl oder einem Forschungsprojekt zu einem bestimmten Zeitpunkt bearbeitet werden und an anderer Stelle und zu anderen Zeitpunkten nur bedingt Interesse an der gleichen Thematik besteht.

Hinzu kommt ein technischer Aspekt, der mit Blick auf die Forschungsgegenstände und die Forschungsmethoden deutlich wird: Qualitative Rezeptionsforschung im Bereich der Medien- und Kommunikationswissenschaft hat in aller Regel die Rezeption von sehr komplexen, multimodalen Medienstimuli im

Blick, seien das Medien-Apps mit Text, Bild, Video und Animationselementen, seien es klassische audiovisuelle Produktionen, wissenschaftliche Präsentationen, Comics oder Infografiken. Schon die Langzeitdokumentation dieser Stimuli an sich kann eine Herausforderung sein – man denke etwa an die Problematik der Sicherung und langfristigen Nutzbarmachung von Medien-Apps heutiger, gängiger Plattformen sowie von Videoformaten oder an den Dokumentationsaufwand, um etwa einen wissenschaftlichen Vortrag mit PowerPoint langfristig intersubjektiv nachvollziehbar zu machen (Video des Vortrags, Folien, eventuell zusätzliches Video zur Dokumentation der Atmosphäre im Vortragsraum). Die Dokumentationsschwierigkeiten sind sogar noch erheblicher, wenn man die üblichen Forschungsmethoden betrachtet: Formen der Befragung, Beobachtungen, mitunter Blickaufzeichnungen. Das Ergebnis des Einsatzes dieser Methoden sind in aller Regel Audio- und/oder Video-Dateien sowie im Falle der Blickaufzeichnung zudem große Datenmengen in Formaten, die nur mittels Spezialsoftware genutzt werden können. Um der Gesellschaft die Ergebnisse solcher Forschung nachvollziehbar zur Verfügung stellen zu können, müsste eine Infrastruktur geschaffen werden, die dazu in der Lage ist, die unterschiedlichen Ergebnisdaten der Forschung und die gesicherten Stimuli präzise miteinander in Bezug zu setzen. Ein Beispiel: Aus der qualitativen Rezeptionsforschung zu wissenschaftlichen Präsentationen mittels PowerPoint weiß man, dass im Falle des Einsatzes einer reinen Textfolie das abschnittsweise Einblenden des Textes zur Steuerung der Rezeption sehr viel besser geeignet ist als das einmalige Einblenden der gesamten Textfolie (vgl. Bucher & Niemann 2012: 293–294). Wollte man den Forschungsprozess, der zu dieser Erkenntnis führte, transparent und intersubjektiv nachvollziehbar mittels eines Dokumentationssystems darlegen, so müsste dieses System folgende Anforderungen erfüllen:

1. Es müsste den Anwendern zu denjenigen Versuchspersonen, die in der entsprechenden Forschung zitiert werden, Videodaten der Vortragsaufzeichnungen vorspielen und
2. Videos der entsprechenden Blickverläufe dieser Personen bereitstellen sowie
3. diese beiden Videoquellen an den relevanten Stellen synchronisieren können.

In anderen Fällen wäre eine solche Materialfülle und Synchronisationsleistung noch mit Interviewdaten aus Audiodateien oder Textdokumentationen zu ergänzen.

Wenn man sich die typischen Forschungsgegenstände der qualitativen Rezeptionsforschung einmal im Detail ansieht, wird schnell deutlich, dass die nachvollziehbare Bereitstellung von Forschungsergebnissen für die Gesellschaft nicht nur auf technischer Ebene zu Herausforderungen führt. Egal ob Nachrichtenbeiträge aus dem Fernsehen, Tweets oder Infografiken in Zeitschriften: Urheber- und Nutzungsrechte machen die schnelle und einfache

Bereitstellung der Forschungsgegenstände, etwa in dem oben skizzierten Dokumentationssystem, sehr schwierig.

Von diesen forschungsstrukturellen, technischen und rechtlichen Aspekten einmal abgesehen, steht die grundsätzliche Frage im Raum, ob die mit einer Standardisierung einhergehenden Vorteile den vermeintlichen zeitlichen und damit auch finanziellen Mehraufwand im Forschungsprozess tatsächlich rechtfertigen. Dahinter steht letztlich die Überlegung, ob die Replizierbarkeit von Forschungsergebnissen, der in der naturwissenschaftlichen und generell in der quantitativen Forschung so große Bedeutung zugemessen wird, in der qualitativen Rezeptionsforschung tatsächlich den gleichen Stellenwert haben muss, wenn die Forschungsergebnisse hier eben nicht den Anspruch eines allgemeingültigen Gesetzes oder gesamtgesellschaftlicher Repräsentativität formulieren.

Im Folgenden soll ein Weg zur Standardisierung der Sicherung und Dokumentation von Forschungsdaten in der qualitativen Rezeptionsforschung beschrieben werden, der den oben dargelegten Einwänden und Schwierigkeiten Rechnung trägt.[1]

4.1 Kleine Forschungseinheiten: Annäherung durch Standardisierung

Die strukturellen Gegebenheiten im Bereich der qualitativen Rezeptionsforschung mögen zwar die Einstiegsschwelle zur Standardisierung vergrößern, sind aber bei genauerer Betrachtung ein zentrales Argument für eine solche Entwicklung: Kleine Forschungsprojekte bzw. Einzelwissenschaftler sind schon aus Gründen der Kapazität nicht in der Lage, Stimuli oder Fragestellungen anderer Akteure im gleichen Feld – etwa in Form replizierender Untersuchungen – in den Blick zu nehmen, wenn die fremden Daten nicht einfach zugänglich sind und nicht in einer Form vorliegen, die mit den eigenen etablierten Arbeitsroutinen kompatibel ist. Insofern können Standardisierungsbemühungen als aussichtsreiche Maßnahme gesehen werden, um die Vernetzung und den Austausch auf der Inhaltsebene innerhalb der qualitativen Rezeptionsforschung nachhaltig zu verbessern und damit die Forschungsqualität zu steigern.

---

1 Da die skizzierten rechtlichen Aspekte ein gesamtgesellschaftliches Problemfeld darstellen, auf dessen Relevanz die Wissenschaft zwar mit Ausdauer und Nachdruck an geeigneter Stelle im politischen System hinweisen kann und sollte, für das sie jedoch durch eigenes Handeln im Forschungsalltag jenseits des schlichten Verzichts der wissenschaftlichen Auseinandersetzung mit einer Vielzahl von medialen Angeboten keine Lösung herbeiführen kann, wird es im Folgenden nicht weiter aufgegriffen.

4.2 Technische Aspekte: Entwicklung in Stufen

Bei der intersubjektiv nachvollziehbaren langfristigen Sicherung von Medienstimuli und Forschungsergebnissen in Form von Audio-, Video- oder Blickdaten ist es sinnvoll, mindestens zwei Nutzergruppen zu differenzieren und in Entwicklungsstufen zu denken. Zunächst ist die Gruppe der Nutzer innerhalb der Scientific Community von der Nutzergruppe „breite Öffentlichkeit“ zu unterscheiden. Bezieht man sich auf erstere und eine eher mittelfristige Zeitperiode, so sind die Sicherungsprobleme bei genauerer Betrachtung durchaus überschaubar: Da jeder Einzelforscher in seinem Bereich auch bisher schon mit entsprechenden Medienstimuli umgeht und einschlägige Methoden der Befragung und Beobachtung nutzt, ist die notwendige Nutzungsexpertise und die entsprechende (Spezial-)Software in der Regel ohnehin vorhanden. Eine Standardisierungslösung müsste daher in puncto Datenspeicherung/Bereitstellung in erster Linie unkompliziert mit Daten in den einschlägigen (Spezial-)Formaten bestückt werden können (etwa per Web-Upload), müsste diese unkompliziert durchsuchbar machen und sie interessierten Forscherkollegen ohne großen Aufwand wieder zur Verfügung stellen können.[2]

Wenn es um die Nutzergruppe „breite Öffentlichkeit“ geht, greifen die oben beschriebenen Schwierigkeiten in vollem Umfang: Hier ist weder von Fachexpertise im Umgang mit Spezialdaten noch von entsprechender technischer Ausstattung auszugehen. Dementsprechend wäre der gesellschaftlichen Forderung nach transparenter und nachvollziehbarer Darstellung von Forschungsergebnissen nur mittels eines neu zu entwickelnden Systems nachzukommen, das in der oben beschriebenen Form in der Lage ist, Forschungsdaten aus unterschiedlichen Quellen in unterschiedlichen Formaten zusammenzuführen und in einer gängigen technischen Umgebung (Beispiel: Webbrowser) ohne Spezialkenntnisse nutzbar zu machen. Ansätze in einer solchen Richtung gab es bereits, beispielsweise im Rahmen des Forschungsprojekts *Interactive Science*, auch wenn der Fokus dort auf der Zusammenführung und Präsentation qualitativer Forschungsdaten innerhalb der Wissenschaft lag (vgl. Web-basiertes System für Präsentationskorpora, Lobin 2009: 162–163). Ein solches System, das den Anforderungen der Nutzergruppe „breite Öffentlichkeit“ gerecht würde, könnte aufgrund der dargelegten Komplexität sicher erst in einer späteren Entwicklungsstufe der Standardisierungsbemühungen zur Verfügung stehen.

2 Der Aspekt der Datendokumentation wurde an dieser Stelle bewusst ausgeklammert und wird separat behandelt.

4.3 Zeitlicher und finanzieller Mehraufwand: Ersparnis auf mittlere Sicht

Klar ist, jede größere Veränderung in etablierten Forschungsabläufen kostet Zeit und damit auch Geld. Das gilt für die hier vorgeschlagenen Standardisierungen bei der Datensicherung und Dokumentation in der qualitativen Rezeptionsforschung wie in jedem anderen Forschungsbereich auch. Betrachtet man den hier im Fokus stehenden Bereich der Forschungstätigkeit genauer, so wird deutlich, dass nach einer kurzen Phase des Mehraufwands schon bald mit erkennbarer Zeit- und somit auch Kostenersparnis zu rechnen ist: Die Sicherung von Medienstimuli und Forschungsdaten findet auch heute schon statt. Veränderungen betreffen hier in erster Linie den Ablauf des Sicherungsprozesses. Bei Einführung eines einheitlichen Systems mit zentralem Speicherort würden in diesem Zusammenhang zwei Problembereiche entfallen, die im Alltag gerade kleiner Forschungseinheiten durchaus virulent sind: Engpässe bei der Speicherkapazität und Auffindbarkeit alter Forschungsdaten (Stichwort DVD-Archiv oder Sammlung externer Festplatten). Im Bereich der Dokumentation von Forschungsergebnissen und der Annotation der Forschungsdaten sieht es zunächst sicher anders aus: Etablierte lehrstuhlspezifische Verfahren müssten durch ein standardisiertes System ersetzt werden, das in der Regel präzisere und umfassendere Dokumentationsleistungen erfordern wird als es vorher der Fall war. Selbst hier ist jedoch fraglich, ob auf mittlere und lange Sicht tatsächlich Mehrkosten entstehen, wenn man bedenkt, dass eine erhöhte Dokumentationsqualität das präzise Auffinden und Zuordnen von eigenen und fremden Forschungsdaten auch noch nach Jahren und in Anwendungskontexten ermöglichen wird, die zum Zeitpunkt der Erhebung vielleicht noch gar nicht auf der Agenda standen.

Wenn Bemühungen zur Standardisierung der Sicherung und Dokumentation von Forschungsdaten in der qualitativen Rezeptionsforschung Erfolg haben sollen, dann muss dies zwangsläufig auch zu Veränderungen an anderen Stellen im Forschungsprozess führen. Ein ganz wesentlicher Aspekt ist in diesem Zusammenhang der Datenschutz[3]: Einverständniserklärungen von Studienteilnehmern müssen dahingehend erweitert werden, dass die erhobenen Daten (z. B. Audio- oder Videodaten, schriftliche Befragungen, Blickdaten) auch jenseits der konkreten Untersuchung in anderen Forschungskontexten und von anderen Personen genutzt werden können. Eventuell muss je nach Studien-

3 Vgl. dazu auch ausführlicher und für die qualitative Sozialforschung allgemein Huschka & Oellers 2013: 12–13.

kontext über eine Differenzierung der langfristigen Nutzungsoptionen nachgedacht werden (andere Forscher vs. breite Öffentlichkeit).

Standardisierung in der Dokumentation von Daten hat auch Auswirkungen auf ihre Erhebung: Variablen, die später zur Nutzbarkeit der Daten zu Vergleichs- und Replikationszwecken genutzt werden sollen – etwa Alter, Geschlecht, Bildungshintergrund etc. – müssen im Studienkontext erhoben werden. Veränderungen sind hier in zwei Bereichen zu sehen. Zum einen führt das Ziel der Nutzung von Studiendaten in anderen Untersuchungskontexten dazu, dass die Zahl der standardisiert zu erhebenden Variablen tendenziell zunimmt. Mag es für den einen Forscher völlig unerheblich sein, über welche Sprachkenntnisse ein Proband verfügt, so ist diese Variable für andere Forscher womöglich ein entscheidendes Untersuchungskriterium. Zum anderen wird die Erhebungsform zumindest im Bereich personenbezogener Variablen ebenfalls eine Standardisierung erfahren müssen, damit etwa Suchroutinen mit archivierten Daten funktionieren. Es wäre also möglich, dass Forscher, die beim Merkmal Geschlecht bisher die Bezeichnungen „m“ und „w“ in ihren Daten genutzt haben, künftig die vereinheitlichenden Bezeichnungen „männlich“ und „weiblich“ nutzen müssten.

Wie bei der Datensicherung selbst werden diese Veränderungen im Forschungsprozess zu Beginn einen zeitlichen und finanziellen Mehraufwand mit sich bringen. Es ist jedoch auch hier davon auszugehen, dass dieser mit zunehmender Routine deutlich abnimmt. Zudem erscheint auch für den Bereich der qualitativen Rezeptionsforschung plausibel, was Heike Solga als eine Konsequenz des Ziels der Datenweitergabe für die qualitative Sozialforschung insgesamt postuliert: Es werde „mit der Beobachtbarkeit und Überprüfbarkeit von Datenerhebungen und Ergebnissen sowohl die Qualität der Datenerhebungen als auch der Datenauswertungen verbessert“ (Solga 2013: 22).

Literatur

Beißwenger, Michael (2007): *Sprachhandlungskoordination in der Chat-Kommunikation*. Berlin, New York: Walter de Gruyter.

Bucher, Hans-Jürgen (2002): Usability – Core feature of interactivity. Empirical results of audience research on internet- and e-business communication. In H. Luczak, A. E. Cakir & G. Cakir (Hrsg.), *Proceedings of the 6th International Scientific Conference on Work with Display Units (WWDU) 2002*, 444–446. Berchtesgarden: World Wide Work.

Bucher, Hans-Jürgen (2009): Das Internet als Netzwerk des Wissens. Zur Dynamik und Qualität von spontanen Wissensordnungen im Web 2.0. In Heiner Fangerau & Thorsten Halling (Hrsg.), *Netzwerke. Allgemeine Theorie oder Universalmetapher in den Wissenschaften? Ein transdisziplinärer Überblick*, 133–171. Bielefeld: transcript Verlag.

Bucher, Hans-Jürgen (2011): „Man sieht, was man hört“ oder: Multimodales Verstehen als interaktionale Aneignung. Eine Blickaufzeichnungsstudie zur audiovisuellen Rezeption. In Jan Schneider & Hartmut Stöckl (Hrsg.), *Medientheorien und Multimodalität. Ein TV-Werbespot – Sieben methodische Beschreibungsansätze*, 109–150. Köln: Herbert von Halem Verlag.

Bucher, Hans-Jürgen (2012): Intermodale Effekte in der audio-visuellen Kommunikation. Blickaufzeichnungsstudie zur Rezeption von zwei Werbespots. In Hans-Jürgen Bucher & Peter Schumacher (Hrsg.), *Interaktionale Rezeptionsforschung. Theorie und Methode der Blickaufzeichnung in der Medienforschung*, 257–296. Wiesbaden: Springer Verlag.

Bucher, Hans-Jürgen (2014): Sprach- und Diskursanalyse in der Medienforschung. In Matthias Karmasin, Matthias Rath & Barbara Thomaß (Hrsg.), *Kommunikationswissenschaft als Integrationsdisziplin*, 271–298. Wiesbaden: Springer VS.

Bucher, Hans-Jürgen (2017): Understanding multimodal meaning making: Theories of multimodality in the light of reception studies. In O. Seizoy & J. Wildfeuer (Hrsg.) (2017, in press), *New studies in multimodality: Conceptual and methodological elaborations*, 91–123. London, New York: Bloomsbury.

Bucher, Hans-Jürgen & Bettina Boy (2018): How informative are information comics? Empirical results from an eye tracking study and knowledge testing. In J. Laubrock, J. Wildfeuer & A. Dunst, *Empirical Comics Research: Digital, Multimodal, and Cognitive Methods*. London, New York: Routledge (erscheint demnächst).

Bucher, Hans-Jürgen, Maria Huggenberger, Martin Sauter & Peter Schumacher (2012): *Publizistische Qualität im lokalen Fernsehen. Eine sendungsbezogene Rezeptionsstudie*, vol. 53. Baden-Baden: Nomos (Angewandte Medienforschung. Schriftenreihe für die Kommunikationswissenschaft).

Bucher, Hans-Juergen & Philipp Niemann (2012): Visualizing science: the reception of Powerpoint presentations. *Visual Communication* 11.3, 283–306.

Bucher, Hans-Jürgen & Philipp Niemann (2015): Medialisierung der Wissenschaftskommunikation: Vom Vortrag zur multimodalen Präsentation. In Mike S. Schäfer, Silje Kristiansen & Heinz Bonfadelli (Hrsg.), *Wissenschaftskommunikation im Wandel*, 68–101. Köln: Herbert von Halem Verlag.

Bucher, Hans-Jürgen, Philipp Niemann & Martin Krieg (2010): Die wissenschaftliche Präsentation als multimodale Kommunikationsform. Empirische Befunde zu Rezeption und Verständlichkeit von Powerpoint-Präsentationen. In Hans-Jürgen Bucher, Thomas Gloning & Kathrin Lehnen (Hrsg.), *Neue Medien – Neue Formate. Ausdifferenzierung und Konvergenz in der Medienkommunikation*, 375–406. Frankfurt a. M.: Campus Verlag.

Bucher, Hans-Jürgen & Peter Schumacher (2006): The relevance of attention for selecting news content. An eye-tracking study on attention patterns in the reception of print- and online media. *Communications. The European Journal of Communications Research* 31.3, 347–368.

Bucher, Hans-Jürgen & Peter Schumacher (Hrsg.) (2012): *Interaktionale Rezeptionsforschung. Theorie und Methode der Blickaufzeichnung in der Medienforschung*. Wiesbaden: Springer Verlag.

Bucher, Hans-Jürgen, Peter Schumacher & Amelie Duckwitz (Hrsg.) (2007): *Mit den Augen der Leser: Broadsheet und Kompakt-Format im Vergleich. Eine Blickaufzeichnungsstudie zur Leser-Blatt-Interaktion*. Darmstadt: Ifra (Ifra Special Report).

Burger, Harald & Martin Luginbühl (2014): *Mediensprache. Eine Einführung in Sprache und Kommunikationsformen der Massenmedien*, 4. neubearbeitete und erweiterte Auflage. Berlin, Boston: de Gruyter.
Dang-Anh, Mark & Jan Oliver Rüdiger (2015): From frequency to sequence: How quantitative methods can inform qualitative analysis of digital media discourse. *10plus1: Living Linguistics 1, Media Linguistics*, 57–73.
Deutsche Forschungsgemeinschaft (DFG) 2016: *Informationsverarbeitung an den Hochschulen – Organisation, Dienste und Systeme. Stellungnahme der Kommission für IT-Infrastruktur für 2011–2020*. Bonn http://www.dfg.de/download/pdf/foerderung/programme/wgi/kfr_stellungnahme_2016_2020.pdf (letzter Zugriff: 6. 11. 2017).
European Strategy Forums on Research Infrastructures (ESFRI) 2016: *Strategy report on research infrastructures. Roadmap 2016*, http://www.esfri.eu/roadmap-2016 (letzter Zugriff: 25. 11. 2017).
Feuerstein, H-J. & H. J. Heringer (1987): Verständlichkeit, Verstehen, Verständnis? Zur Kommunikation zwischen Fachleuten und Laien. In L. Bress (Hrsg.), *Medizin und Gesellschaft. Ethik – Ökonomie – Ökologie*, 175–188. Berlin, Heidelberg: Springer-Verlag.
Fritz, Gerd & Erich Straßner (1996): *Die Sprache der ersten deutschen Wochenzeitungen im 17. Jahrhundert*, vol. 41. Tübingen: Niemeyer Verlag, (Medien in Forschung und Unterricht).
Gehl, Dagmar (2013): *Vom Betrachten zum Verstehen. Die Diagnose von Rezeptionsprozessen und Wissensveränderungen bei multimodalen Printclustern*. Wiesbaden: Springer.
GESIS – Leibniz-Institut für Sozialwissenschaften e. V. (o. J.): *Datenarchivierung, Internetseite mit Inforationen zur Tätigkeit des GESIS im Bereich der Archivierung von Daten der quantitativen Sozialforschung*. http://www.gesis.org/angebot/archivieren-und-registrieren/datenarchivierung/ (letzter Zugriff: 20. 3. 2017).
Huschka, Denis & Claudia Oellers (2013): Einführung: Warum qualitative Daten und ihre Sekundäranalyse wichtig sind. In Denis Huschka et al. (Hrsg.), *Forschungsinfrastrukturen für die qualitative Sozialforschung*, 9–16. Berlin: SCIVERO.
Leonhard, Joachim-Felix, Hans-Werner Ludwig, Dietrich Schwarze & Erich Straßner (2002): *Medienwissenschaft. Ein Handbuch zur Entwicklung der Medien und Kommunikationsformen*, vol. 15.3. Berlin, New York: Walter de Gruyter.
Lobin, Henning (2009): *Inszeniertes Reden auf der Medienbühne. Zur Linguistik und Rhetorik der wissenschaftlichen Präsentation*. Frankfurt a. M.: Campus.
Muckenhaupt, M. (1980): Der Ärger mit Wörtern und Bildern. Probleme der Verständlichkeit und des Zusammenhangs von Text und Bild. *Kodikas/Code. An intern. Journal of Sem* 2, 187–209.
Schumacher, Peter (2009): *Rezeption als Interaktion. Wahrnehmung und Nutzung multimodaler Darstellungsformen im Online-Journalismus, Internet Research edn*. Baden-Baden: Nomos.
Schröder, Thomas (1995): *Die ersten Zeitungen. Textgestaltung und Nachrichtenauswahl*. Tübingen: Gunter Narr Verlag.
Solga, Heike (2013): Was lässt sich von den Forschungsinfrastrukturen in der empirischen Sozial- und Wirtschaftswissenschaft für qualitative Forschungsinfrastrukturen lernen? In Denis Huschka et al. (Hrsg.), *Forschungsinfrastrukturen für die qualitative Sozialforschung*, 19–24. Berlin: SCIVERO.
Spillner, Bernd (1995): *Sprache. Verstehen und Verständlichkeit*. Frankfurt a. M.: Peter Lang Verlag (Kongreßbeiträge zur Gesellschaft für angewandte Linguistik: Kongreßbeiträge zur Jahrestagung der Gesellschaft für Angewandte Linguistik).

Thimm, Caja, Jessica Einspänner & Mark Dang-Anh (2012): Twitter als Wahlkampfmedium. Modellierung und Analyse politischer Social Media Nutzung. *Publizistik* 57.3, 293–313.

Wissenschaftsrat (2011): *Empfehlungen zu Forschungsinfrastrukturen in den Geistes- und Sozialwissenschaften*. Berlin (28. 1. 2011). https://www.wissenschaftsrat.de/download/archiv/10465–11.pdf (letzter Zugriff: 16. 11. 2017)

III Korpora und Informationsysteme

Ruxandra Cosma und Marc Kupietz

9 Von Schienen, Zügen und linguistischen Fragestellungen

Abstract: Das hier vorgeführte Schienenbild ist das in Anlehnung an Wittenburg (2009) als Erweiterungsinstrument gewählte Mittel in dem Versuch, Computertechnologie, linguistische Forschung und Vernetzung am Institut für Deutsche Sprache in deren rasch wachsenden Vielschichtigkeit zu beschreiben. Hier werden u. a. drei Blickwinkel, der des Technologie entwickelnden Wissenschaftlers, des entwickelnden Nutzers und des Nutzers von Informationstechnologie in der linguistischen Forschung vereint und um eine für den Sprachvergleich neue Dimension, die sprachspezifische Parameter von Analyseinstrumenten miteinander harmonisiert, erweitert.

Keywords: Empirie, Forschungswerkzeuge, Sprachkorpora, Vernetzung

1 Die Reise hin

In der Beobachtung sprachlicher Daten und Phänomene, deren Untersuchung, Beschreibung und Erklärung fährt man bereits seit Jahrzehnten auf linearen Transportbahnen, die sich immer mehr vernetzen. In kurzen Wegen zwischen einzelnen Knotenpunkten werden entfernteste Ziele miteinander verbunden. Die Fahrt erfolgt innerhalb der eigenen Sprache, überschreitet auch Grenzen, richtet sich sprachen- und netzvergleichend aus. Manchmal trifft man auf ein ähnlich ausgebautes Schienennetz, manchmal muss man sich die Reise anders überlegen, oder auch einen Gleisanschluss organisieren. Die Schienen zum *spurengebundenen Verkehr* (z. B. Ross 2001: 8) erlauben in der Linguistik mittlerweile staunenerregende Fahrgeschwindigkeit. Die komplementären Seiten des sich auf Forschung begebenden Nutzers und des Infrastruktur betreibenden, in Stand haltenden und selbst erstellenden Unternehmers führen in der germanistischen Linguistik-Forschung zusammen, anders als im Schienenverkehr,

Ruxandra Cosma, Universität Bukarest, Fakultät für Fremdsprachen, Abteilung für Germanische Sprachen, Str. Pitar Mos 7–13, RO-010451 Bucuresti, E-Mail: ruxandra.cosma@lls.unibuc.ro
Marc Kupietz, Institut für Deutsche Sprache, R5 6–13, D-68161 Mannheim,
E-Mail: kupietz@ids-mannheim.de

https://doi.org/10.1515/9783110538663-010

wo Transport zur Raumüberwindung und Schieneninfrastruktur in Eisenbahnunternehmen organisatorisch getrennt werden.[1]

Digitalisierung und Vernetzung verändern auf breiter Front die Welt, das Denken, die Wissenschaft. Digitalisierung und Vernetzung, sowie sich daraus ergebende Expansion und Diversifikation bestimmen auch das linguistische Arbeitsfeld. Im Kontext der in diesem Aufsatz gewählten Konzepte entfalten sich in dem linguistischen Bereich immer raffiniertere Fragen zu Datenerhebung, zu ihrer Verfügbarmachung, zum Umgang mit Daten, zu Datenmengen, zu Methoden und Interpretation, zur Nutzung von bestehenden technologischen Diensten, zu technischen Herausforderungen. Diese zunehmende Komplexität, im Sinne der Begriffsverfeinerung und trennung, zeigt, wie bedeutend Aufbau, Ausbau und Nutzung von Infrastruktur geworden sind. Eine die Linguistik begleitende Forschung, mit Hilfe derer schon lange bestehende linguistische Fragestellungen neu betrachtet und behandelt werden können, macht die Nutzung der Infrastruktur zum Bestandteil der linguistischen Wissenschaft und zur linguistischen Arbeit selbst. Aus der Nutzerperspektive sind Daten, Methoden, Technik und Dienste Teil der Forschungsinfrastruktur, wie sie aber auch Teil der eigenen wissenschaftlichen Auseinandersetzung mit Sprache sind.

Nach wie vor ist linguistische Forschung ohne institutionelle Infrastruktur möglich; der Sprachwissenschaftler kann selbst Daten sammeln, analysieren, interpretieren, Phänomene erklären. Voraussetzungen, Verfahrensweisen, Interessen, Blickrichtungen, Zielsetzungen, Grundbegriffe sind individuell begründet. Für die datenbasierte, empirische Linguistik sind aber die Ansprüche an die Validität und Nachvollziehbarkeit der Prinzipien der Theoriebildung sowie der Methoden und empirischen Verfahren und die Exaktheit ihrer Beschreibung, also letztlich an Wissenschaftlichkeit, erheblich gestiegen. Empirisch und theoretisch ausgerichtet, ist Linguistik heutzutage meist viel mehr als nur Erfahrungswissenschaft und theoretische Wissenschaft, die „nicht nur die sprachliche Materialbasis, sondern auch die Reflexion auf die die Materialbasis erfassenden Methoden zum Objekt ihrer Forschung erhebt“ (van de Velde 1970: 7). Computertechnologie ergänzt und unterstützt sie nun, und diese ist aus der linguistischen Forschung in den auf authentische Daten angewiesenen Bereichen kaum noch wegzudenken, wie sie in Ergänzung dann auch auf quantitative Methoden der Datenanalyse, die deskriptive Statistik oder statistische Operationalisierung nebst qualitativen Untersuchungen immer mehr zurückgreifen muss. Diese methodischen Erweiterungen erweisen sich als notwendig, da die Erkenntnisse einer Korpusuntersuchung die Frage nach der Generalisierbarkeit

1 Vgl. Ross (2001: 8–9).

der Aussagen zum untersuchten Phänomen überstehen sowie eventuell anstehende Zweifel ausräumen müssen. Damit hängt die Wahl und Beschreibung des Sprachdaten-Ausschnitts zusammen. Und dies hängt seinerseits von der Größe des Korpus ab. Somit kommt man zurück auf das Thema der Methodenentwicklung und auf Wissenschaftlichkeitskriterien, auf Fragen nach der wissenschaftlichen Objektivität und der Repräsentativität, die nicht absolut beschrieben werden können. Die Technikdistanz, von Jakob (1991) für die Geisteswissenschaften noch als deutlich lang beschrieben, wobei Technik den Bereich der Kommunikation und Information miteinbezog, ist in der Linguistik allgemein, besonders aber in der Korpuslinguistik so viel kürzer geworden. So auch die nicht nur dadurch schrumpfende Distanz zu den *sciences*, zu den ‚harten' Wissenschaften (vgl. Busse 1989; Teubert & Belica 2014: 298). Denn Korpuslinguistik setzt in diesem Sinne viel mehr als nur Informationstechnologie ein, sie modelliert auch sprachliche Daten und Sachverhalte zu verschiedenen Verwendungszwecken (z. B. Jannidis, Kohle & Rehbein 2017: 13–14, 100–104).

Die Schienen-Analogie, die wir hier als Ausgangspunkt angenommen haben und zu eigenen Interessen ausbauen, geht auf Peter Wittenburg (2009: 7) zurück, der sie gerne zur Illustration des Potenzials der Vernetzung bestehender Forschungsressourcen verwendet hat. Für Wittenburg sind solche Schienen (und Signalanlagen) im Vergleich zu den komplexen Zügen, die in seiner Analogie Methoden und Werkzeuge darstellen, technisch betrachtet äußerst simpel und haben dennoch das Potenzial, die Produktivität von Sprachforschern deutlich zu verbessern.

Für den Sprachwissenschaftler und dessen Auseinandersetzung mit sprachlichen Phänomenen sind vor allem Reiseziele und der Weg hin von Bedeutung; diese mögen, je nach Vorgehensart, mit der Nutzung einer Infrastruktur, als Bestandteil ihrer Wissenschaft, einhergehen. Ist einem Reisenden das Reiseziel bekannt, so sucht dieser nach Reiserouten, gegebenenfalls nach Zugverbindungen, die zum Ziel führen sollen. Er vergewissert sich, ob das Ziel über Schienenverkehr erreichbar ist, informiert sich somit darüber, wie die bestehende Infrastruktur für eigene Ziele optimal genutzt werden kann. Auf der Reise zum Ziel fährt man durch die Landschaft, direkt oder über Verkehrsknoten, die ein Umsteigen möglich machen. Dadurch wird das bestehende Netzwerk noch intensiver genutzt. Es mag durchaus sein, dass sich der hierfür gewählte Vergleich und das In-Beziehung-Setzen (Hentschel 2010: 15–21) von Schienenverkehr und textbasierter Technologie zur Analyse von Sprache nicht gänzlich als mängelfrei erweist, dass im Kleinen noch Ungereimtheiten zu erkennen sind. Die Ähnlichkeitsbetrachtung ist hier aber breit angelegt und verfolgt keine disparaten Eigenschaften, sondern strukturelle Ähnlichkeiten und

Ähnlichkeiten zwischen Relationen.[2] In diesem Fall zwischen Objekt, Fragestellungen, Instrument und Methode.

Die bildliche Klammerung zum wissenschaftsergänzenden Unterbau bietet die Möglichkeit, Vernetzung sinnfällig zu machen. Nicht nur als Bild von Knoten und Kanten und vom kürzesten Weg, als Vernetzungstopologie, sondern auch im Sinne des Verbundes mit europäischen Projekten.

Am Institut für Deutsche Sprache (IDS) stehen Entwickler und Nutzer von Infrastruktur in engster Zusammenarbeit bzw. sind in der wissenschaftlichen Arbeit auf verschiedenste Weise vereint. Der vorliegende Aufsatz beschreibt einige dieser vielen Komponenten, die dem Aufbau und Nutzung von Sprachdaten und Analyseinstrumenten dienen. Heterogene Anforderungen an digitale Infrastrukturen im Dienste der Linguistik, wie die Frage nach der Zusammenbringung von Daten, Methoden, Diensten etc., um Forschungsaufgaben effektiver lösen zu können, oder die Frage, wie man Daten, Methoden etc. möglichst nachhaltig nutzbar machen kann, beschreiben einen Aufbruch, der am IDS schon lange Tradition hat.

Im Folgenden werden wir uns auf Datenerfassung, Modellierungsprozesse und Auswertungsverfahren zur geschriebenen Sprache am IDS beziehen. Die am IDS parallel verlaufenden Entwicklungen zu Korpora gesprochener Sprache werden z. B. in Stift & Schmidt (2014) beschrieben und in Kupietz & Schmidt (2015) zu denen der Korpora geschriebener Sprache in Beziehung gesetzt.

2 Digitalisierung

Die Dokumentation sprachlicher Phänomene am Korpus, in den ersten Monaten nach der Gründung des Instituts 1964 bereits als Aufgabenbereich des IDS umrissen, daraufhin Datenerhebung, d. h. Texterfassung und Kodierung zur grammatischen Analyse[3] oder zu lexikografischen Zwecken mittels datenverarbeitender Maschinen, haben sich im Wandel der Zeit und über die Zeit hinweg etabliert. Über die – selbst für die heutige Zeit über-ambitionierten – anfänglichen Erwartungen schreibt Engel (1968: 1), dass in der Zeit nirgends so viele Blütenträume zerstört worden seien wie auf dem Gebiet der maschinellen Sprachverarbeitung.

2 „Analogy denotes a resemblance not between things, but between the relations of things.“ (Hentschel 2010: 24).

3 Vorerst im Rahmen des IDS-Projekts zu Grundstrukturen der deutschen Sprache (siehe Teubert & Belica 2014: 301).

Die Anfänge wandten sich der Datenerfassung zur Erstellung des Mannheimer Korpus I und II, oder einer Datenbank zum Wortschatz von Ost und West als Zeitungskorpus zu. Zunächst durch Einsatz von Schreiblochern und Lochstreifen mittels weniger Programme zur Kodierung und Datenverarbeitung im Rahmen einer Kooperation mit Rechenzentren in Darmstadt oder in Bonn, Ende der 1960er Jahre bereits im Rahmen eines eigenen Rechenzentrums am Institut über Programme, die von wissenschaftlichen Mitarbeitern der linguistischen Datenverarbeitung erstellt wurden, ferner stark erweitert durch das 1998 gestartete Verbundprojekt Deutsches Referenzkorpus (DeReKo-I) (Teubert & Belica 2014: 308) wuchsen die Datenbanken über Zeit hinweg zu einer universellen Ur-Stichprobe des geschriebenen Gegenwartsdeutsch, dem Deutschen Referenzkorpus DeReKo (Kupietz et al. 2008). Die Programme zur Auswertung der Daten wurden ihrerseits zu immer komplexeren Analysesystemen, mit denen immer komplexere linguistische Ziele verfolgt werden konnten, entwickelt; hier seien Programme und Übergänge vorerst nur kurz erwähnt – REFER, COSMAS I, COSMAS II, KorAP. Das dynamische DeReKo (Kupietz & Keibel 2009a) bot von vornherein die Möglichkeit zur Erstellung eines eigenen (virtuellen) Korpus, das individuellen Forschungsinteressen und Untersuchungsintentionen entgegenkommt, indem sich für die speziellen Fragestellungen und zu untersuchende Grundgesamtheiten maßgeschneiderte und möglichst repräsentative Unterstichproben approximieren lassen. Dokumentiert werden diese Entwicklungen in ihren Anfängen und in ihrem zeitlichen Verlauf von Engel (1968, 2014), Zint (1968), Stickel (2007, 2014), Berens (2014), Hellmann (2014), Kupietz (2014), Teubert und Belica (2014) und vielen anderen. Den Übergang von der maschinellen zur elektronischen und digitalen Datenerfassung, wählt man differenzierende Bezeichnungen, unterstützten technische und institutionelle Rahmenbedingungen, die bereits sehr früh vorwegnahmen, dass die Zukunft einem Schienenverkehr und der Vernetzung, anfangs nur als Computernetz betrachtet, gehört. Anfänge der Vernetzungsprozesse zwischen den Ressourcen oder auch zwischen Personen, die Wandlung von Informationstechnologien zu Informations- und Kommunikationstechnologien sind in der Mitte der 1980er Jahre angesiedelt (Jannidis, Hubertus & Rehbein 2017: 9). Ein institutionalisierter Raum für die Grundlagenforschung zur linguistischen Methodik basierend auf Korpora, d. h. für einen eigenen Programmbereich Korpuslinguistik, wurde am IDS ab 2004 geboten (Kupietz 2014, Teubert & Belica 2014: 315). 2009 ging dann aus diesem, angestoßen durch das BMBF-geförderte Projekt *Aufbau eines Zentrums „Digitale Forschungsressourcen für die germanistische Sprachwissenschaft“* ein eigener Programmbereich Forschungsinfrastrukturen hervor, der zunächst vor allem Informationsangebote, Informationstechnik und Informationswissenschaft, Projekte „mit Schnittstellen-

charakter“ wie CLARIN-D, WissGrid, TextGrid etc. sowie juristische Expertise bündelte (Schonefeld & Witt 2014).

Jeder der Aspekte des medialen Transfers von gedruckten Texten in eine elektronische Fassung, zusammenhängend mit sich rasch und selbstständig entwickelnden Rekonstruktions- und Kodierungsfragen, Fragen zum Umgang mit *Big Data* mittels eines Analysesystems, bereitete eine nächste Phase vor: die der Vernetzung, nicht nur im Sinne des *global village*, das durch das immer noch verbesserbare Internet und somit durch den Online-Zugriff möglich wird, sondern im Sinne einer Sammlung von Ressourcen, die den Traum von einem vereinfachten gebündelten Zugriff auf diese Ressourcen erlaubt. Und hier wird wiederum die eingangs genannte Analogie zu den linearen Transportbahnen merklich: Durch die Landschaft der Forschungsdaten fährt man mit ICE-Zügen, die wie im eigentlichen Sinne spur- und schienengebundene Verkehrsmittel sind. In diesem Sinne seien hier die am IDS entwickelten und eingesetzten Auswertungssysteme COSMAS II und KorAP vorerst nur kurz genannt; deren Rolle wird im Folgenden auf einer allgemeineren Ebene diskutiert.

3 Wissenschaftliche Werkzeuge

Inzwischen steht schon längst fest, dass „Korpuslinguistik nicht auf eine Senkung der theoretischen, sondern auf eine Hebung der methodischen Ansprüche hinausläuft“ (Lehmann 2007: 27). Damit hängt jedoch die noch nicht geklärte Frage nach der Rolle der Infrastruktur zusammen, ob diese nur ein Instrument im Dienste der Linguistik darstellt, oder vielmehr ein elementarer Bestandteil der eigentlichen Forschung wie die Interpretation von Daten und Analyseergebnissen ist. Zur Beantwortung der Frage ist zunächst vorauszuschicken, dass die Unterscheidung zwischen Infrastruktur und Forschung sowie zwischen Infrastruktur und Wissenschaft meist vom jeweiligen Betrachtungswinkel und -abstand abhängt.[4] So kann der Large-Hadron-Collider (LHC) mit etwas Abstand sicherlich als Bestandteil der CERN-Infrastruktur oder besser, da diese zur Forschung verwendet wird, als Bestandteil der CERN-Forschungsinfrastruktur betrachtet werden. Auf der anderen Seite sind Teilchenbeschleuniger im Allgemeinen natürlich Gegenstand von Forschung und der LHC selbst auch Ergebnis von Forschung. Zwischen Forschung und Infrastruktur(-Bestandteil) muss also grundsätzlich kein Widerspruch bestehen.

4 Eine weit gefasste Definition des Begriffs Forschungsinfrastruktur findet sich in Wissenschaftsrat 2011: 17.

Interessanter ist vielleicht die Frage, inwieweit solche wissenschaftlichen Werkzeuge, die der Produktion oder der Auswertung von Forschungsdaten dienen, derselben fachwissenschaftlichen Forschung zuzurechnen sind. Zwar kann auch diese Frage einfach darauf zurückgeführt werden, wie eng man diese Kategorie fasst, jedoch hat es diesbezüglich in vielen Disziplinen eine rasante Entwicklung gegeben. Dies betrifft generell solche Disziplinen, in denen anhand von großen Mengen an Forschungsdaten statistische Modelle gewonnen werden können, die konventionelle Erklärungsmodelle zunehmend ergänzen oder gar ersetzen, also keineswegs nur die Teilchenphysik, sondern auch z. B. auch die Medizin und solche Geisteswissenschaften, die auf größere Datenmengen zurückgreifen können und deren Gegenstand zu komplex ist oder sich aus anderen Gründen (noch) einer *Theory of Everything* verschließen (siehe z. B. Lehmann 2007, Keibel & Kupietz 2009). In besonderem Maße ist dabei die Linguistik betroffen, bei deren Forschungsprimärdaten vom Typ „beobachtbare Sprache" es sich gewissermaßen um ein bewegliches Ziel handelt, das zumindest Aspekte eines dynamischen Artefakts aufweist, da es sich zwar eingeschränkt aber permanent und unvorhersagbar verändert (vgl. Keller 1994; Kupietz & Keibel 2009b). Hinzu kommt, dass man von diesen und anderen Forschungsdatentypen in der Linguistik nur sehr indirekt auf allgemeinere Prinzipien – seien sie sprachsystemischer, sprachpsychologischer oder anderer Art – schließen kann, um zu neuen Hypothesen und Erkenntnisgewinn zu gelangen. Wie ein solcher Schluss gelingen kann, welche Eigenschaften der Daten dabei relevant und welche theoretischen Konstrukte dabei hilfreich sind, ist daher selbst elementarer Gegenstand der sprachwissenschaftlichen Forschung. Dies manifestiert sich z. B. in der großen Zahl unterschiedlicher Grammatiktheorien und -formalismen in der Linguistik, die sich mit unterschiedlichen Blickwinkeln ihrem Gegenstand nähern, parallel verfolgt werden und deren Anzahl die in der Teilchenphysik bei weitem übersteigt. Einen Grund für die Annahme, dass dieser Theorienpluralismus sich durch die verstärkte Einbeziehung von Korpora und anderen Daten, vielleicht durch eine Art Objektivierung, zwangsläufig wesentlich reduzieren würde, gibt es kaum. Es ist lediglich zu erwarten und bereits zu beobachten, dass Theorien, die sich überhaupt nicht anhand von Daten falsifizieren lassen, es schwer haben werden. Die Bandbreite möglicher Ansätze – und entsprechend die Unsicherheit über den jeweils richtigen – beginnt auch noch in einer eng gefassten Korpuslinguistik bereits bei der Auswahl und Erhebung der Daten. Sie setzt sich fort z. B. bei der Kodierung der Daten bzw. der Kodierung der Rekonstruktion der ursprünglichen Sprachereignisse, über die verwendeten Analysemethoden (jeweils mit unterschiedlichem Skopus, unterschiedlichen Bewertungsmaßen und anderen Parametern) bis hin zur Darstellung der Such- oder Analyseergebnisse. Für die

Ergebnisse einer abschließenden Interpretation haben alle auf diesem Weg getroffenen Entscheidungen einen potenziell ausschlaggebenden Einfluss. Daten und Werkzeuge zu ihrer Aufbereitung, Analyse und Interpretation ausschließlich unter dem Aspekt einer nicht weiter aufschlüsselbaren und einer idealerweise austauschbaren Infrastruktur zu betrachten, birgt daher die Gefahr, die Relevanz eines ganzen Bereichs ausschlaggebender Entscheidungen zu unterschätzen. Dies würde nicht nur die Suche nach möglichen Fehlschlüssen unnötig einschränken, sondern generell den Suchraum für Wege zum Erkenntnisgewinn.

Die Beantwortung der Frage, ob man diesen Bereich innerhalb einer jeweiligen Fachwissenschaft verorten sollte, hängt davon ab, in welchem Maße ein Werkzeug für die jeweilige Fachwissenschaft konzipiert wurde. Zwar kann es sein, dass auch eher fachfremde Werkzeuge sich als nützlich erweisen (hier besteht eher das Problem nur scheinbare Nützlichkeit zu erkennen). Dass aber etwa speziell linguistische Werkzeuge nur zufällig nützlich für die linguistische Forschung sind, ist eher unwahrscheinlich. Vielmehr muss man davon ausgehen, dass die oben genannte Reihe von Entscheidungen jeweils mit dem Ziel getroffen wurden, linguistische Probleme lösen zu können. Hinzu kommt, dass durch den Einzug von Korpora in verschiedene Bereiche der Linguistik, in denen sie sehr unterschiedliche Rollen spielen können, derzeit viele noch wenig erforschte Stellen gibt, an denen quantitative auf qualitative Methoden aufeinandertreffen. Diese Stellen betreffen Analysewerkzeuge nicht nur hinsichtlich ihrer verwendeten statistischen Methoden, sondern z. B. auch hinsichtlich der Darstellung, d. h. Visualisierung der Ergebnisse dieser, da sie einen erheblichen Einfluss auf die aus ihnen abduzierbaren qualitativen Hypothesen haben.

Die Frage, ob eine Darstellung sinnvoll ist, lässt sich auch hier nicht allgemein mit ja oder nein beantworten. Vielmehr ist die Antwort typischerweise abhängig vom konkreten Forschungsziel. Sie ist also mehr als nur disziplinspezifisch. Ein weiterer Aspekt der Integration von quantitativen und qualitativen Methoden betrifft die Unterstützung von Arbeitsabläufen durch Analysewerkzeuge. Hier haben Analysewerkzeuge ein großes Potenzial, quantitative Methoden mit qualitativen Methoden zu engen Zyklen miteinander zu verzahnen und in explorativen Kontexten die Abduktion von Hypothesen stark zu beschleunigen (vgl. Kupietz & Schmidt 2015: 307; Kupietz et al. 2017a: 326 f.). An dieser Stelle können verlässliche wissenschaftliche Werkzeuge auch dazu beitragen, dass in der Linguistik auch einzelne Wissenschaftler weiterhin empirisch fundierte Forschung zumindest zu einem großen Teil eigenständig betreiben können. Eine zunehmende Spezialisierung und Arbeitsteilung, wie sie z. B. in der Psychologie (s. a. Haider 2015) schon lange gang und gäbe ist, ist zwar aufgrund des Wachstums des notwendigen Wissens unumgänglich, aber

zumindest begrenzt auf die Exploration von Daten und Hypothesen können Werkzeuge z. B. Experten zur Datenanalyse ersetzen.

Wittenburgs klare Unterscheidung zwischen Schienen und (ICE-)Zügen ist also nicht nur hilfreich, um das Potenzial einer genuin infrastrukturellen und technisch vergleichsweise simplen Vernetzung zu verdeutlichen, sondern auch, um mögliche Missverständnisse bezüglich der Rolle wissenschaftlicher Werkzeuge zu vermeiden.

4 Vernetzung

Bereits in den 1990er Jahren gab es zahlreiche paneuropäische Projekte zur Förderung von *human language technology* (HLT), zum Aufbau von Ressourcen und Technologien sowie zum Aufbau von Infrastrukturen, wie insbesondere das Projekt TELRI (*Trans-European Language Resources Infrastructure*) (siehe Teubert & Belica 2014). Zwar haben sich diese Initiativen positiv auf die Zusammenarbeit europäischer Sprachzentren und auf die Entwicklung der Korpuslinguistik auch außerhalb der Anglistik und über die Lexikologie hinaus ausgewirkt, aber letztlich „wenige Ressourcen von bleibendem Wert" (Teubert & Belica 2014: 306) und insbesondere wenig greifbare Strukturen oder Infrastrukturen von bleibendem Wert geschaffen. Sprachressourcen und Technologien wurden vielmehr außerhalb solcher koordinierender Verbundinitiativen und im Zuge zunehmender Diversifikation und Spezialisierung auseinander divergierender Subdisziplinen entwickelt.

Die Ausgangslage für die nächste große paneuropäische Infrastrukturinitiative CLARIN (*Common Language Resources and Technology Infrastructure*) war 2006 folglich, dass durchaus in großer Anzahl vorhandene Sprachressourcen und Technologien (Language Resources and Technology – LRT) zum weitaus größten Teil untereinander nicht interoperabel waren. Das heißt, es war nicht ohne weiteres möglich, ein Korpus vom Zentrum A mit einem Korpus von Zentrum B zu vereinen, beide zusammen mit einem Wortart-Tagger vom Zentrum C zu annotieren, um dann die neu entstandene Ressource mit einem Werkzeug vom Zentrum D zu analysieren. Um dies zu bewerkstelligen, mussten nach dem sogenannten „*download first, then process*"-Paradigma (Kemps-Snijders et al. 2008) zunächst alle Ressourcen und Werkzeuge ggf. lizenziert und heruntergeladen werden, um dann die Auszeichnungsformate der Ressourcen und die Ein- und Ausgabeformate der Werkzeuge aneinander anzupassen und einen größten gemeinsamen Nenner aller Lizenzbedingungen zu finden. Der damit verbundene Aufwand wäre allerdings so groß gewesen, dass er in der Regel unrealistisch und höchstens teilweise zu bewerkstelligen war, so

dass es für solche Projekte meist einfacher war, neue, spezialisierte Ressourcen aufzubauen und Werkzeuge zu entwickeln, die allerdings selbst wiederum typischerweise nicht in anderen Kontexten nachnutzbar waren.

Die Hoffnung von CLARIN war, dass sich durch eine Verbesserung der Interoperabilität bestehender und zukünftiger LRT ihre Wiederverwendbarkeit exponentiell steigern ließe und damit ein regelrechter Boost in den mit Sprachressourcen arbeitenden geistes- und sozialwissenschaftlichen Disziplinen ausgelöst würde. Um eine solche Interoperabilitätsverbesserung zu erreichen, hat CLARIN vor allem an den folgenden Teilen einer besseren und stabileren Vernetzung von Zentren, Ressourcen und Nutzern gearbeitet[5]:

1. Sozialer Teil
   a) Interoperabilität als Gegenstand von Kooperationen (über unmittelbare fachliche Ziele hinaus)
   b) Erweiterung der Kontakte von Diensteanbietern über ihre unmittelbaren Zielgruppen hinaus, durch so genannte Facharbeitsgruppen zu verschiedenen Themenbereichen und durch Kurationsprojekte
   c) Dissemination z. B. durch Sommerschulen und durch Help-Desks
2. Normativer Teil
   a) Standards und Best-Practices zu Auszeichnungs- und Kodierungsformaten
   b) Best-Practices z. B. zum Aufbau und zur Kuration von Korpora
   c) Schnittstellendefinitionen
   d) Muster-Lizenzvereinbarungen
3. Rechtlich/vertraglicher Teil
   a) z. B. Verträge zwischen Diensteanbietern
4. Informationstechnischer Teil
   a) Basisdienste wie PID-Service und AAI
   b) Muster-Applikationen

Hinsichtlich einer Basisinfrastruktur hat CLARIN nun zehn Jahre später in der Tat sehr gute Dienste geleistet. Für die allermeisten Anwendungen liegt zwar ein web-basiertes Paradigma, wie es noch in Kemps-Snijders et al. (2008) skizziert wird, bei dem Ressourcen und Tools per Mausklick virtuell kombiniert werden können, in weiter Ferne, aber es sind die Grundlagen dafür geschaffen, dass auch im Rahmen von Abschlussarbeiten und kleinen Projekten Daten und Methoden aufgebaut, genutzt und nachnutzbar gemacht werden können. Der Grund, warum die Vision eines Mausklick-Szenarios nicht erreicht wurde und

5 Eine detaillierte Darstellung der verschiedenen Infrastrukturkomponenten findet sich in Fiedler et al. 2014.

wahrscheinlich auch nicht erreicht werden kann, ist dabei einfach die bei der Entwicklung von Ressourcen und Technologien mit begrenzten Mitteln unabdingbare Verpflichtung zur Sparsamkeit. Dies hat zur Folge, dass Features, die für die primäre Anwendung nicht notwendig sind gegenüber Features, die für die primäre Anwendung wichtiger sind, zurückstehen müssen, so dass sich LRT typischerweise nur bedingt und nicht auf dem gleichen Niveau in nicht primär intendierten Anwendungskontexten wiederverwenden lassen. Im CLARIN-Kontext trifft dies in besonderem Maße zu, da der CLARIN-Skopus nicht an bestimmten Anwendungen orientiert ist, sondern am Datentyp Sprachressource im Allgemeinen, so dass das Anwendungsspektrum innerhalb der Sozial- und Geisteswissenschaften sehr breit ist und völlig unterschiedliche Forschungsgegenstände und -ziele umfasst, obwohl mit DARIAH[6] zeitgleich eine weitere, komplementäre Infrastruktur für die Geistes- und Kulturwissenschaft aufgebaut wurde. Dass CLARIN trotzdem viel erreicht hat, liegt vor allem daran, dass sich die Arbeiten auf solche basalen Infrastrukturschichten konzentriert haben, die zwar für den Anwender nicht unmittelbar sichtbar sind, aber tatsächlich praktisch immer benötigt werden, um Sprachressourcen für die eigene Forschung zu erschließen und die Ergebnisse für andere nachnutzbar zu machen. So hat CLARIN ein recht vollständiges infrastrukturelles Basisfundament geschaffen, auf dem ein breites Spektrum fachwissenschaftlich orientierter Projekte aufsetzen kann.

Dass CLARIN sich auf, aus Anwenderperspektive meist weit entfernte, basale Komponenten konzentriert hat, war auch richtig und unvermeidbar im Hinblick auf eine andere für Infrastrukturen unabdingbare Eigenschaft: ihre Nachhaltigkeit. Wenn nicht sichergestellt ist, dass Schienen noch in vielen Jahren benutzbar sind, ist es ein Fehler, in ihren Ausbau oder ihre langfristige Nutzung zu investieren. Diese Herausforderung betrifft in erster Linie die technischen Infrastrukturkomponenten, deren permanente Instandhaltung und Wartung dauerhaft gesichert sein muss. Die Frage was und wie viel gerade an technischer Infrastruktur entwickelt werden sollte – also wie viel eine Disziplin sich leisten kann und leisten sollte, ist demnach weniger von den initialen Entwicklungskosten abhängig als vielmehr von der Frage, wie viel dauerhaft finanzierbar ist. Auch wenn die Infrastruktur-Finanzierung in den Geisteswissenschaften noch nicht so stark bedarfsgekoppelt ist wie in den Naturwissenschaften, müssen die Kosten natürlich auch hier in angemessener Relation zur Größe der Anwenderschaft stehen, so dass auch aus diesem Grund einfache Basisdienste, angefangen bei der Vergabe persistenter Identifikatoren

6 https://de.dariah.eu/ (letzter Zugriff: 20.10.2017).

bis hin vielleicht zu einer einfachen föderierten Suche (Stehouwer et al. 2012) im Fokus stehen müssen.

Durch eigene Initiativen wird der Vernetzungsgedanke hausintern weiterentwickelt, das IDS wächst durch Ausbau zur Mitte. Einige dieser Initiativen werden in den folgenden Abschnitten beschrieben.

5 EuReCo

Die Idee der EuReCo-Initiative (siehe Kupietz et al. 2017b), die 2013 im Kontext von CLARIN und im Kontext der CMLC-Workshops (Bański et al. 2013, 2015; Kupietz et al. 2014a) entstanden ist, besteht darin, ein virtuelles Europäisches Referenzkorpus zu schaffen, das ausschließlich auf bereits existierenden oder im Aufbau befindlichen Referenz- und Nationalkorpora beruht, und zu diesen einen gemeinsamen Zugang anzubieten. EuReCo verfolgt dabei im Wesentlichen zwei Ziele: Zum einen soll so mit Hilfe der gemeinsamen technischen Plattform sprachvergleichende Forschung mit dynamisch definierbaren, vergleichbaren Korpora der verschiedenen Sprachen ermöglicht werden. Zum anderen soll durch die Verwendung einer gemeinsamen Plattform eine neue bzw. erweiterte Infrastruktur geschaffen werden, die auf folgende Synergieeffekte abzielt:

1. Vermeidung von Redundanzen bei der Entwicklung von Korpusanalyse- und Rechercheplattformen: Funktionalitäten werden nur einmal entwickelt und in verschiedenen Kontexten genutzt.
2. Vermeidung von Redundanzen bei Aufbau und Kuration cross-linguistischer und vergleichbarer Korpora: Bestehende Korpora werden dynamisch nachgenutzt. Sie werden weiterhin von den jeweils verantwortlichen Institutionen erweitert und kuratiert und über EuReCo nur weiteren Verwendungsarten zugeführt.
3. Vermeidung redundanter Lizenzierung: Die einzelsprachlichen Korpora verbleiben an ihren Standorten und müssen daher für die neuen Verwendungsarten typischerweise nicht neu lizenziert werden.
4. Verbesserte Nachhaltigkeit insbesondere der technischen Infrastrukturelemente durch die Unterstützung mehrerer Institutionen.
5. Schnellerer methodischer Fortschritt in den einzelsprachlichen Philologien durch die gemeinsame Verwendung von Weiterentwicklung von Analysemethoden und Workflows.
6. Leichtere Identifikation zusätzlich möglicher Synergien durch engere Kooperation unter den National- und Referenzkorpusanbietern.

Das oben skizzierte Problem, dass Forschungsinfrastrukturen typischerweise entweder sehr basal sind oder nicht mächtig genug, um stark divergierenden Endnutzeranforderungen zu genügen, soll dabei gelöst werden, indem die technische Infrastruktur als Ergänzung und zusätzliche Alternative zur möglicherweise bereits bestehenden Infrastruktur verwendet wird. Darüber hinaus soll der technische Teil der Infrastruktur als Open-Source-Software mit Entwicklungszweigen realisiert sein, die, falls notwendig, voneinander abweichen können.

5.1 DRuKoLA

EuReCo existierte lange Zeit nur als Idee und als Teil verschiedener Förderanträge. Dies änderte sich Anfang 2016 mit dem Start des Projektes DRuKoLA (ursprünglich *Sprachvergleich korpustechnologisch – Deutsch-Rumänisch* genannt), einer Kooperation zwischen dem IDS, der Universität Bukarest und den Forschungsinstituten der Rumänischen Akademie in Bukarest und Iași.[7] In seinem Mittelpunkt stehen das Deutsche Referenzkorpus (DeReKo) (Kupietz et al. 2010) und das Referenzkorpus der Rumänischen Gegenwartssprache (*Reference Corpus of Contemporary Romanian Language*, CoRoLa) (Tufiș et al. 2015, 2016). DRuKoLA ist ein transdisziplinäres Projekt mit Zielen im Bereich der Korpuslinguistik, der Grammatik und der Forschungsinfrastrukturentwicklung, indem über die Entwicklung und Harmonisierung von Forschungsinfrastruktur auf Ziele in der Grammatik und in der Korpuslinguistik hingearbeitet wird. Aufgrund neu entwickelter Instrumente sollen probeweise explorative linguistische Studien durchgeführt werden Somit wird eine Grundlage für empirische und technische Forschung im Sprachvergleich (hier Deutsch-Rumänisch) geschaffen, die sich gleichwohl einem höheren Ziel, der Entwicklung einer EuReCo-Grundstruktur und einer Blaupause zur Integration weiterer Korpora bzw. weiterer Institutionen in die Infrastruktur, unterwirft.

Neben dem ersten Schritt zur Verwirklichung der Idee eines virtuellen Europäischen Referenzkorpus ist also der Sprachvergleich das zentrale Thema des DRuKoLA-Projekts. Während im Wechsel der Zeiten der Sprachvergleich in Verbindung zur Sprachtechnologie im Bereich automatischer Übersetzung, der Erstellung von zwei- oder mehrsprachigen elektronischen Wörterbüchern oder

---

7 DRuKoLA wird von der Alexander-von-Humboldt-Stiftung als Institutspartnerschaft zwischen dem IDS, der Universität Bukarest und der beiden Forschungsinstitute für Computertechnologie der rumänischen Akademie in Bukarest und in Iași (Mihai Drăgănescu Research Institute for Artificial Intelligence in Bukarest und Institute for Computer Science in Iași) gefördert.

von multilingualen und multifunktionalen Sprachdatenbanken (z. B. Rovere & Wotjak 1993: 2) angesiedelt war, verbindet das IDS über DRuKoLA dessen lange Tradition des Sprachvergleichs mit der Vorreiter-Rolle im Bereich der Gewinnung und Auswertung von Forschungsdaten zur deutschen Sprache.

Am Institut für Deutsche Sprache wurde dem Deutschen im Sprachvergleich schon immer ein besonderer Stellenwert zugemessen. In unterschiedlicher Weise war das Deutsche Gegenstand sprachvergleichender grammatischer Forschungsprojekte, sei es in Form von zweisprachig kontrastiven Grammatiken, über das den sprachlichen Nachbarraum des Deutschen untersuchende GDE-Projekt (Projekt *Grammatik des Deutschen im europäischen Vergleich*) mit der 2017 fertiggestellten *Grammatik des Deutschen im europäischen Vergleich. Das Nominal* (Gunkel et al. 2017), gegenwärtig fortgesetzt mit dem Thema *Verbgrammatik* (u. a. Trawiński 2016; Wöllstein 2016), über mehrsprachige Forschungsprojekte wie das *Eurogr@mm-Projekt* (z. B. Augustin & Fabricius-Hansen 2012; Dalmas, Fabricius-Hansen & Schwinn 2016), über die Erforschung komplexer Sätze im Deutschen, Portugiesischen und Italienischen (u. a. Blühdorn & Ravetto 2014) oder komplexer Argumentstrukturen im Deutschen und im Rumänischen (Cosma et al. 2014), um hier nur einige Projekte der jüngsten Zeit zu nennen. Dabei zeigt sich immer wieder, dass eine sprachübergreifende Perspektive auf die untersuchten Phänomene das Sichtfeld ergänzt, so wie Analyseansätze zu grammatischen Phänomenen in anderen Sprachen neue Einblicke für die Analyse entsprechender Phänomene in der eigenen Sprache bringen können. Um mit den Worten der Autoren des GDE-Projekts den wissenschaftlichen Ertrag des Sprachvergleichs zu beschreiben, werde dadurch ersichtlich, „wo eine Sprache eigene Wege geht und wo ihre Strukturen mit denen anderer Sprachen konvergieren“ (Gunkel & Zifonun 2012, Klappentext).

Jedoch sind die sich immer komplexer gestaltenden anwendungsbezogenen Ziele sprachvergleichender Untersuchungen nicht zu übersehen. Am Beispiel von DRuKoLA als wegeröffnenden Versuch des anspruchsvollen hausinternen Zieles, das neue Analysesystem KorAP „für die Allgemeinheit zu öffnen“ und „auf mehrere Schultern zu verteilen“ (Kupietz 2014: 326), daher Ressourcen aus unterschiedlichen Sprachen unter einem System zwecks Weiterentwicklung zu bündeln, wird vielmehr gezeigt, dass im Sinne der gleichen, in diesem Aufsatz verwendeten Analogie zum Ziel des Sprachvergleichs, aber auch über den Sprachvergleich hinaus, aus einer technologischen Perspektive Gleisanschlüsse organisiert werden können, um Knotenpunkte grenzübergreifend miteinander zu verbinden. Im Falle des als Vergleichssprache gewählten Rumänischen fällt der Gleisanschluss mit der Erstellung des rumänischen Referenzkorpus CoRoLa zusammen, einem Projekt, mit dem ein Jahr vor dem Start des DruKoLA-Projekts begonnen wurde und das ein Jahr vor diesem enden

soll, daher optimale Bedingungen für Harmonisierungsverfahren zur Daten- und Prozessmodellierung bietet. Zum anderen aber bietet die Möglichkeit, mit demselben Analysesystem zwei Sprachen z. B. syntaktisch anhand virtuell definierbarer Korpora zu untersuchen, ein herausragendes Beispiel für die Art, in der jeweils einzelsprachliche Untersuchungen zum selben Phänomen über die gleiche Ausgangsbasis und dieselben technologischen Möglichkeiten im Sprachvergleich sprachtechnologische und infrastrukturelle Entwicklungen begünstigen können, die ihrerseits wiederum u. a. zur Weiterentwicklung der Linguistik im engeren Sinne beitragen. Als nächste Sprache soll das Ungarische im Rahmen des Projekts *DeutUng*[8] in EuReCo integriert werden.

5.2 KorAP

Die technische Basis des DRuKoLA-Projekts bildet die Korpusanalyseplattform KorAP (Kupietz & Frick 2013; Bański et al. 2013; Diewald et al. 2016), die als Nachfolgesystem zum Korpusrecherche- und Analysesystem COSMAS II (Bodmer 2005; Bodmer Mory 2014) entwickelt wird. KorAP ist zwar ein in erster Linie wissenschaftliches Werkzeug und bietet in dieser Hinsicht einige Neuerungen (Kupietz et al. 2017a), es bietet aber auch unter dem infrastrukturellen Aspekt Designeigenschaften, die für die Verwirklichung der DRuKoLA-Projektziele und der EuReCo-Visionen unabdingbar sind. Die in dieser Hinsicht wichtigste Eigenschaft ist, dass KorAP es erlaubt, mit Daten zu arbeiten, die physikalisch auf verschiedene Standorte verteilt sind. Dies setzt nicht nur das EuReCo-Konzept um, dass jedes beteiligte Zentrum weiterhin für die Kuration seiner Korpora verantwortlich ist, sondern löst auch das in der Linguistik omnipräsente rechtlich-ökonomische Problem, dass Korpustexte normalerweise nur so weit lizenziert werden können, dass sie zwar auszugsweise, aber nicht vollständig die Standorte der unmittelbaren Lizenznehmer verlassen dürfen (Kupietz et al. 2014b: 4; Schonefeld & Witt 2014: 332). Ein weiteres damit eng verbundenes Problem ist, dass die (Nach-)Nutzbarkeit von Daten und Recherchesoftware in Zusammenhang mit extern entwickelten Analysemethoden typischerweise sehr eingeschränkt ist. KorAPs Lösungsansatz folgt hier Jim Grays (2003) Postulat *put the computation near the data*. D. h., wenn die Daten aus rechtlichen oder informatischen Gründen nicht bewegt werden können, müssen Möglichkeiten geschaffen werden, dass vielmehr die Methoden zu den Daten bewegt und auf diese angewendet werden können. Nach der ursprüng-

---

8 Langtitel: Deutsch-ungarischer Sprachvergleich: korpustechnologisch, funktional-semantisch und sprachdidaktisch – ebenfalls als Institutspartnerschaft (zwischen der Universität Szeged und dem IDS) gefördert von der Alexander von Humboldt-Stiftung.

lichen KorAP-Konzeption sollte dieses Prinzip durch eine spezielle, gesicherte Umgebung für extern entwickelte Methoden (*mobile code sandbox)* realisiert werden (vgl. Kupietz 2014: 325). Dieser Ansatz wurde jedoch zugunsten einer besseren langfristigen Wartbarkeit durch einen einfacheren, auf Standardtechniken der Open-Source-Softwareentwicklung beruhenden Ansatz ersetzt. So können neue Algorithmen an den bestehenden KorAP-Schnittstellen ansetzen, die z. B. auch von KorAPs Weboberfläche verwendet werden. Falls dies nicht möglich ist, können die neuen Methoden aber auch als KorAP-Erweiterungen realisiert werden, die entweder in eigenen Entwicklungszweigen gepflegt werden oder, idealerweise, in den Hauptentwicklungszweig von KorAP eingehen. Bei aller gebotenen Skepsis bezüglich der Nachnutzbarkeit von spezialisierten Forschungswerkzeugen in neuen Kontexten, legt KorAP so zumindest eine auch aufgrund des institutionellen Engagements des IDS realistische infrastrukturelle Basis für weitergehende externe Entwicklungen. Inwieweit diese genutzt wird und so vielleicht auch zur Bündelung von Ressourcen und zur Kanonisierung von Methoden sowie damit vielleicht auch zur Beantwortung der Fragen, wie viel Infrastruktur sich die Linguistik leisten kann und leisten sollte, beitragen kann, bleibt allerdings abzuwarten.

6 Rundblick

Im Unterschied zum wirklichen Schienenverkehr, in dem Streckenausbau mit Fahrplan-Umstellungen einhergeht[9], gibt es in der hier beschriebenen Entwicklung von Forschungsinfrastruktur in der linguistischen Forschung am IDS trotz Verkürzung der Fahrten über Schienenwege und Höchstgeschwindigkeitszüge keinen Zeitzwang. Dafür gibt es aber immer mehr neue Schienen, Gleise, Anschlüsse, Knotenpunkte, Weichen und Übergänge, die von dem Stellwerk am IDS entwickelt, durchgeführt, angeboten werden. Das Netzwerk wächst in dem inneren Bau weiter, öffnet sich nach außen, schließt sich an. Der Weg hin verläuft aber in eigenem Tempo.

Der vorliegende Text berichtet letztendlich von Erfahrungen, von Ergebnissen, von Träumen und Neuerungen. Wir übernehmen hier den Werbeslogan des neuen Porsche-Panamera: *Ideen alleine verändern nichts. Sondern der Mut, sie umzusetzen.*[10] Hiermit sei für den Mut zum Wandel gedankt.

9 Z. B. https://web.de/magazine/reise/streckenausbau-deutsche-bahn-plant-grosse-fahrplan-aenderungen-32381716 (letzter Zugriff: 20. 10. 2017).

10 http://www.porsche-club-deutschland.de/PcLife/16-2/PC-Life/PDFs/032_Porsche.pdf (letzter Zugriff: 20. 10. 2017).

Literatur

Augustin, Hagen & Cathrine Fabricius-Hansen (Hrsg.) (2012): *Flexionsmorphologie des Deutschen aus kontrastiver Sicht*. Tübingen: Groos.

Bański, Piotr, Joachim Bingel, Nils Diewald, Elena Frick, Michael Hanl, Marc Kupietz, Piotr Pęzik, Carsten Schnober & Andreas Witt (2013): KorAP: The new corpus analysis platform at IDS Mannheim. In: Vetulani, Zygmunt & Hans Uszkoreit (Hrsg.): *Human language technologies as a challenge for computer science and linguistics*. Proceedings of the 6th Language and Technology Conference, 586–587. Poznan: Uniwersytet im. Adama Mickiewicza w Poznaniu. http://nbn-resolving.de/urn:nbn:de:bsz:mh39-32617 (letzter Zugriff: 20. 10. 2017).

Bański, Piotr, Hanno Biber, Evelyn Breiteneder, Marc Kupietz, Harald Lüngen, Andreas Witt (Hrsg.) (2015): *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3)*, Lancaster, 20 July 2015. Mannheim: Institut für Deutsche Sprache.

Berens, Franz Josef (2014): Zur Frühgeschichte des IDS. In *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, 52–60. Mannheim: Institut für Deutsche Sprache.

Blühdorn, Hardarik & Miriam Ravetto (2014): Satzstruktur und adverbiale Subordination. Eine Studie zum Deutschen und zum Italienischen. In *Linguistik online* 67, 3–44.

Busse, Dietrich (1989): Sprachwissenschaftliche Terminologie Verständlichkeits- und Vermittlungsprobleme der linguistischen Fachsprache. In *Muttersprache* Jg. 99, 27–38.

Bodmer, Franck (2005): COSMAS II. Recherchieren in den Korpora des IDS. In *Sprachreport* 3, 2–5.

Bodmer Mory, Franck (2014): Mit COSMAS II „in den Weiten der IDS-Korpora unterwegs“. In *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, 376–385. Mannheim: Institut für Deutsche Sprache.

Cosma, Ruxandra, Stefan Engelberg, Susan Schlotthauer, Speranța Stănescu & Gisela Zifonun (Hrsg.) (2014): *Komplexe Argumentstrukturen. Kontrastive Untersuchungen zum Deutschen, Rumänischen und Englischen*. Berlin, München, Boston: de Gruyter. (Konvergenz und Divergenz 3).

Dalmas, Martine, Cathrine Fabricius-Hansen & Horst Schwinn (Hrsg.) (2016): *Variation im europäischen Kontrast. Untersuchungen zum Satzanfang im Deutschen, Französischen, Norwegischen, Polnischen und Ungarischen*. Berlin, Boston: de Gruyter. (Konvergenz und Divergenz 5).

Diewald, Nils, Michael Hanl, Eliza Margaretha, Joachim Bingel, Marc Kupietz, Piotr Bański & Andreas Witt (2016): KorAP architecture – Diving in the deep sea of corpus data. In Nicoletta Calzolari et al. (Hrsg.): *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 3586–3591. Portorož, Slovenia, Paris: European Language Resources Association (ELRA).

Engel, Ulrich (1968): Vorbemerkung zum Forschungsbericht des Instituts für deutsche Sprache. In Hugo Moser, Peter von Polenz, Hans Glinz & Paul Grebe (Hrsg.): *Forschungsberichte des Instituts für deutsche Sprache*. Mannheim: Institut für deutsche Sprache.

Engel, Ulrich (2014): Das Institut für deutsche Sprache 1965–1976. In *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, 64–79. Mannheim: Institut für Deutsche Sprache.

Fiedler, Norman, Antonina Werthmann, Maik Stührenberg, Oliver Schonefeld, Joachim Bingel, Andreas Witt (2014): *Forschungsinfrastrukturen in außeruniversitären Forschungseinrichtungen*. Forschungsbericht. Mannheim: Institut für Deutsche Sprache.

Gray, Jim (2003): *Distributed computing economics*. Technical Report MSR-TR-2003–24, Microsoft Research. https://www.microsoft.com/en-us/research/publication/distributed-computing-economics/ (letzter Zugriff: 24.11. 2017).

Gunkel, Lutz, Adriano Murelli, Susan Schlotthauer, Bernd Wiese & Gisela Zifonun (2017): *Grammatik des Deutschen im Europäischen Vergleich. Das Nominal*. Berlin, Boston: de Gruyter. (Schriften des Instituts für Deutsche Sprache 14).

Gunkel, Lutz & Gisela Zifonun (Hrsg.) (2012): *Deutsch im Sprachvergleich. Grammatische Kontraste und Konvergenzen*. Institut für Deutsche Sprache. Jahrbuch 2011. Berlin, Boston: de Gruyter.

Haider, Hubert (2015): *Zwischen Sortieren und Prädizieren – Grammatikforschung & Grammatikbeschreibung auf dem Sprung vom 19. in das 21. Jahrhundert.* Vortrag im Rahmen der Tagung *ars grammatica*. Institut für Deutsche Sprache, Mannheim. Abstract: http://arsgrammatica.ids-mannheim.de/abstracts/Abstract_Haider.pdf (letzter Zugriff: 20.10. 2017).

Hellmann, Manfred W. (2014): Die Bonner „Forschungsstelle für öffentlichen Sprachgebrauch“ (F.ö.S.) 1964–1980. In *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache. Mannheim*. 80–98. Mannheim: Institut für Deutsche Sprache.

Hentschel, Klaus (2010): Die Funktion von Analogien in den Naturwissenschaften. In *Acta Historica Leopoldina* Nr. 56, 13–66.

Jakob, Karlheinz (1991): *Maschine, mentales Modell, Metapher: Studien zur Semantik und Geschichte der Techniksprache*. Tübingen: Niemeyer.

Jannidis, Fotis, Hubertus Kohle & Malte Rehbein (Hrsg.) (2017): *Digital Humanities. Eine Einführung*. Stuttgart: J. B. Metzler.

Keibel, Holger & Marc Kupietz (2009): Approaching grammar: Towards an empirical linguistic research programme. In Makoto Minegishi & Yuji Kawaguchi (Hrsg.): *Working Papers in Corpus-based Linguistics and Language Education*, No. 3, 61–76. Tokyo: Tokyo University of Foreign Studies.

Kemps-Snijders, Marc, A. Klassmann, C. Zinn, P. Berck, A. Russel & Peter Wittenburg (2008): Exploring and enriching a language resource archive via the web. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 807–810. Paris: European Language Resources Association (ELRA).

Kupietz, Marc & Holger Keibel (2009a): The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In Makoto Minegishi & Yuji Kawaguchi (Hrsg.), *Working Papers in Corpus-based Linguistics and Language Education*, No. 3, 53–59. Tokyo: Tokyo University of Foreign Studies.

Kupietz, Marc & Holger Keibel (2009b): Gebrauchsbasierte Grammatik: Statistische Regelhaftigkeit. In Marek Konopka & Bruno Strecker (Hrsg.), *Deutsche Grammatik – Regeln, Normen, Sprachgebrauch*, 33–50. Berlin, New York: de Gruyter. (= Jahrbuch des Instituts für deutsche Sprache 2008).

Kupietz, Marc & Elena Frick (2013): Korpusanalyseplattform der nächsten Generation. In Iva Kratochvílová & Norbert Richard Wolf (Hrsg.), *Grundlagen einer sprachwissenschaftlichen Quellenkunde*, 27–36. Tübingen: Narr, 2013. (= Studien zur Deutschen Sprache 66).

Kupietz, Marc (2014): Der Programmbereich Korpuslinguistik am IDS. In: *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*. Mannheim. 320–328.

Kupietz, Marc, Hanno Biber, Harald Lüngen, Piotr Bański, Evelyn Breiteneder, Karlheinz Mörth, Andreas Witt & Jani Takhsha (Hrsg.) (2014a): *Proceedings of the LREC 2014 Workshop Challenges in the Management of Large Corpora (CMLC-2)*. Paris: European Language Resources Association (ELRA).

Kupietz, Marc, Harald Lüngen, Piotr Bański, Cyril Belica (2014b): Maximizing the potential of very large corpora. In Marc Kupietz et al. (Hrsg.), *Proceedings of the LREC-2014-Workshop Challenges in the Management of Large Corpora (CMLC-2)*, 1–6. Paris: European Language Resources Association (ELRA).

Kupietz, Marc & Thomas Schmidt (2015): Schriftliche und mündliche Korpora am IDS als Grundlage für die empirische Forschung. In Ludwig M. Eichinger (Hrsg.), *Sprachwissenschaft im Fokus. Positionsbestimmungen und Perspektiven*, 297–322. Berlin/Boston: de Gruyter. (Jahrbuch des Instituts für Deutsche Sprache 2014).

Kupietz, Marc, Nils Diewald, Michael Hanl, Eliza Margaretha (2017a): Möglichkeiten der Erforschung grammatischer Variation mithilfe von KorAP. In Marek Konopka & Angelika Wöllstein (Hrsg.), *Grammatische Variation. Empirische Zugänge und theoretische Modellierung*, 319–329. Berlin, Boston: de Gruyter. (Jahrbuch des Instituts für Deutsche Sprache 2016).

Kupietz, Marc, Andreas Witt, Piotr Bański, Dan Tufiș, Dan Cristea & Tamás Váradi (2017b): EuReCo – Joining forces for a European Reference Corpus as a sustainable base for cross-linguistic research. In *Proceedings of the 5th Workshop on Challenges in the Management of Large Corpora (CMLC-5)*, 15–19. Mannheim: Institut für Deutsche Sprache.

Lehmann, Christian (2007): Daten – Korpora – Dokumentation. In Werner Kallmeyer & Gisela Zifonun (Hrsg.), *Sprachkorpora – Datenmengen und Erkenntnisfortschritt*. Berlin, New York: de Gruyter (Jahrbuch des Instituts für Deutsche Sprache 2006) http://www.christianlehmann.eu/publ/daten.pdf (letzter Zugriff: 20. 10. 2017).

Ross, Sebastian (2001): *Strategische Infrastrukturplanung im Schienenverkehr. Entwicklung eines Planungs- und Entscheidungsmodells für die deutsche Bahn AG*. Diss. Deutscher Universitätsverlag. Wiesbaden: Gabler.

Rovere, Giovanni & Gerd Wotjak (Hrsg.) (1993): *Studien zum romanisch-deutschen Sprachvergleich*. Tübingen: Niemeyer. (Linguistische Arbeiten 297).

Schonefeld, Oliver & Andreas Witt (2014): Forschungsinfrastrukturen am IDS: Gegenwart und Zukunft. In *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, 329–336. Mannheim: Institut für Deutsche Sprache.

Stehouwer, Herman, Matej Durco, Eric Auer & Daan Broeder (2012): Federated search: Towards a common search infrastructure. In N. Calzolari et al. (Hrsg.), *Proceedings of LREC 2012: 8th International Conference on Language Resources and Evaluation*, 3255–3259. Paris: European Language Resources Association (ELRA).

Stickel, Gerhard (2007): Die Gründerjahre des IDS. In Heidrun Kämper & Ludwig M. Eichinger (Hrsg.), *Sprach-Perspektiven. Germanistische Linguistik und das Institut für Deutsche Sprache*, 23–41. Tübingen: Narr.

Stickel, Gerhard (2014): Schlechte und bessere Zeiten für das IDS. In *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, 122–143. Mannheim: Institut für Deutsche Sprache.

Stift, Ulf-Michael & Thomas Schmidt (2014): Mündliche Korpora am IDS: Vom Deutschen Spracharchiv zur Datenbank für Gesprochenes Deutsch. In *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, 360–375. Mannheim: Institut für Deutsche Sprache.

Teubert, Wolfgang & Cyril Belica (2014): Von der linguistischen Datenverarbeitung am IDS zur „Mannheimer Schule der Korpuslinguistik“. In: *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, 298–319. Mannheim: Institut für Deutsche Sprache.

Trawiński, Beata (2016): Linguistic data in contrastive studies. Addressing the need for a multilingual parallel resource annotated with semantic-functional information, In Domínguez Vázquez, María José & Silvia Kutscher (Hrsg.), *Interacción entre gramática, didáctica y lexicografía. Estudios contrastivos y multicontrastivos*, 85–98. Berlin, Boston: De Gruyter.

Tufiș, Dan, Verginica Barbu Mititelu, Elena Irimia, Ștefan D. Dumitrescu, Tiberiu Boroș, Horia N. Teodorescu, Dan Cristea, Andrei Scutelnicu, Cecilia Bolea, Alex Moruz & Laura Pistol (2015): CoRoLa starts blooming – An update on the Reference Corpus of Contemporary Romanian Language. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3)*, 5–10.

Tufiș, Dan, Verginica Barbu Mititelu, Elena Irimia, Ștefan, D. Dumitrescu, Tiberiu Boroș (2016): The IPR-cleared Corpus of Contemporary Written and Spoken Romanian Language. In Nicoletta Calzolari et al. (Hrsg.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16*), Portoroz, Slovenia.

Velde, Roger G. van de (1970): Zur Wissenschaftlichkeit der Linguistik. *Studia philosophica Gandensia* vol. 8, 7–33.

Wissenschaftsrat (2011) *Übergreifende Empfehlungen zu Informationsinfrastrukturen*, Berlin. http://www.wissenschaftsrat.de/download/archiv/10466-11.pdf (letzter Zugriff: 20. 10. 2017).

Wittenburg, Peter (2009): *Service centers as backbone of a D-SPIN infrastructure – requirements and current state*. Vortrag vom 17. Oktober 2009 im Rahmen des D-SPIN-Beiratstreffens. https://weblicht.sfs.uni-tuebingen.de/Beirat/2-MPI-service-centers.pdf (letzter Zugriff: 20. 10. 2017).

Wöllstein, Angelika (2016): Was ein Strukturmodell für die (kontrastive) Sprachbetrachtung im DaF-Bereich leisten kann. In Juan Cuartero Otal, Larreta Zulategui, Juan Pablo & Christoph Ehlers (Hrsg.): *Querschnitt durch die deutsche Sprache aus spanischer Sicht. Perspektiven der Kontrastiven Linguistik*, 211–231. Berlin: Frank & Timme.

Alexander Geyken, Matthias Boenig, Susanne Haaf,
Bryan Jurish, Christian Thomas und Frank Wiegand

10 Das Deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN

Abstract: Im Zentrum dieses Beitrags steht das *Deutsche Textarchiv*, eine Plattform zum Korpusaufbau und zur Korpusanalyse, die im Kontext aller geistes- und sozialwissenschaftlichen Disziplinen mit historischen Fragestellungen nutzbar ist. Das DTA, das am Zentrum Sprache der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) angesiedelt ist, wurde von 2007 bis 2016 von der Deutschen Forschungsgemeinschaft (DFG) gefördert und bildet mittlerweile eine wesentliche Komponente der Forschungsdateninfrastruktur des deutschen Teils von CLARIN. Der Beitrag präsentiert das DTA als webbasierte Forschungsplattform sowohl für die Erstellung und die Kuration von Korpustexten als auch für die Korpusanalyse und verortet es innerhalb der (digitalen) Geisteswissenschaften.

Keywords: Annotation, Historische Fragestellungen, Qualitätssicherung, Sprachkorpora, XML

1 Einführung

Ziel des *Deutschen Textarchivs* (DTA)[1] ist die Erstellung eines disziplinen- und gattungsübergreifenden Grundbestands deutschsprachiger Texte aus dem Zeitraum von ca. 1600 bis etwa 1900. Die Textauswahl erfolgte auf der Grundlage einer von Akademiemitgliedern der BBAW kommentierten und ergänzten, umfangreichen Bibliographie. Aus dieser wurde von der DTA-Projektgruppe ein nach Textsorten und Disziplinen ausgewogenes Textkorpus zusammengestellt, das als Grundlage für ein Referenzkorpus zur Entwicklung der neuhoch-

1 http://www.deutschestextarchiv.de (letzter Zugriff: 6. 11. 2017).

Alexander Geyken, Matthias Boenig, Susanne Haaf, Bryan Jurish, Christian Thomas, Frank Wiegand, Berlin-Brandenburgische Akademie der Wissenschaften, Jägerstraße 22/23, D-10117 Berlin, E-Mail: dta@bbaw.de

https://doi.org/10.1515/9783110538663-011

deutschen Sprache dient. Um den historischen Sprachstand möglichst genau abzubilden, wurden als Vorlage für die Digitalisierung in der Regel die Erstausgaben der Werke zugrunde gelegt. Das nach diesen Kriterien zusammengestellte DTA-Kernkorpus wird kontinuierlich erweitert. Derzeit umfasst es etwa 1.500 Werke mit einem Umfang von etwa 120 Millionen Textwörtern (Stand April 2017).

Neben seiner Funktion als Korpusaufbauprojekt wurde das DTA von Beginn an auch als aktives Archiv konzipiert. Es soll Ort der Anlagerung weiterer Korpora sein. Hierzu wurden zunächst Qualitätskriterien bezüglich der Metadaten, der Auszeichnungstiefe der Textstrukturen sowie der Genauigkeit des Volltexts festgelegt. Des Weiteren wurde aus der Vielzahl der verschiedenen Texte ein für viele Kontexte nutzbares Strukturformat entwickelt, das neben seiner Funktion als Austauschformat für verschiedene Korpora die Interoperabilität für so verschiedene Anwendungsfälle wie die Korpusanzeige, die Volltextsuche und das Textmining gewährleistet. Mit dem DTA-Basisformat (DTABf) liegt ein solches Format vor, welches mit dem XML/TEI-P5-Standard vollständig kompatibel ist und das mittlerweile auch eine weit über das DTA hinausgehende Verbreitung gefunden hat (dazu mehr in Abschnitt 2.1). Zur Qualitätssicherung der Volltexte und der Strukturdaten wurde darüber hinaus mit DTAQ eine webbasierte Plattform entwickelt, die das verteilte Korrekturlesen und Korrigieren von Texten erlaubt. Hierzu wurden flexible Möglichkeiten zum Textimport aus unterschiedlichen Formaten und eine Text-Bild-Ansicht geschaffen sowie ein Editor in die Plattform integriert, mit dem Texte ohne zusätzlich zu installierende Software bearbeitet werden können (siehe dazu Abschnitt 2.2–4). Am Ende des Korrekturprozesses steht die Veröffentlichung auf der DTA-Webseite, wo die Werke über eine Text-Bild-Ansicht zugänglich sowie an verschiedene Analysewerkzeuge angebunden sind. Letztere führten dazu, dass sich das DTA mit DTAQ von einer Korrektur- und Veröffentlichungsplattform zu einer Forschungsplattform entwickelte. Mit CAB (*Cascaded Analysis Broker*; vgl. dazu Jurish 2012), einem Werkzeug zur Normalisierung historischer Schreibweisen, wird eine schreibweisentolerante Volltextsuche über alle Texte des DTA bereitgestellt. Mit der Integration von GermaNet (Hamp & Feldweg 1997; Henrich & Hinrichs 2010), einer lexikalischen Ressource, die Substantive, Verben und Adjektive nach Bedeutungsähnlichkeit in SynSets zusammenfasst, wird darüber hinaus auch die Volltextsuche nach semantischen Kategorien ermöglicht. Ferner stehen eine Reihe von lexikometrischen Analysewerkzeugen zur Verfügung, insbesondere zu zeitlichen Verläufen von Wortfrequenzen, zu diachronen Kollokationen sowie eine auf den Voyant-Tools basierende quantitative Textanalyse (siehe hierzu Abschnitt 3).

Die Integration von Texten in das DTA stellt keine Einbahnstraße dar: Alle Texte des DTA stehen unter einer offenen Lizenz und können damit ohne

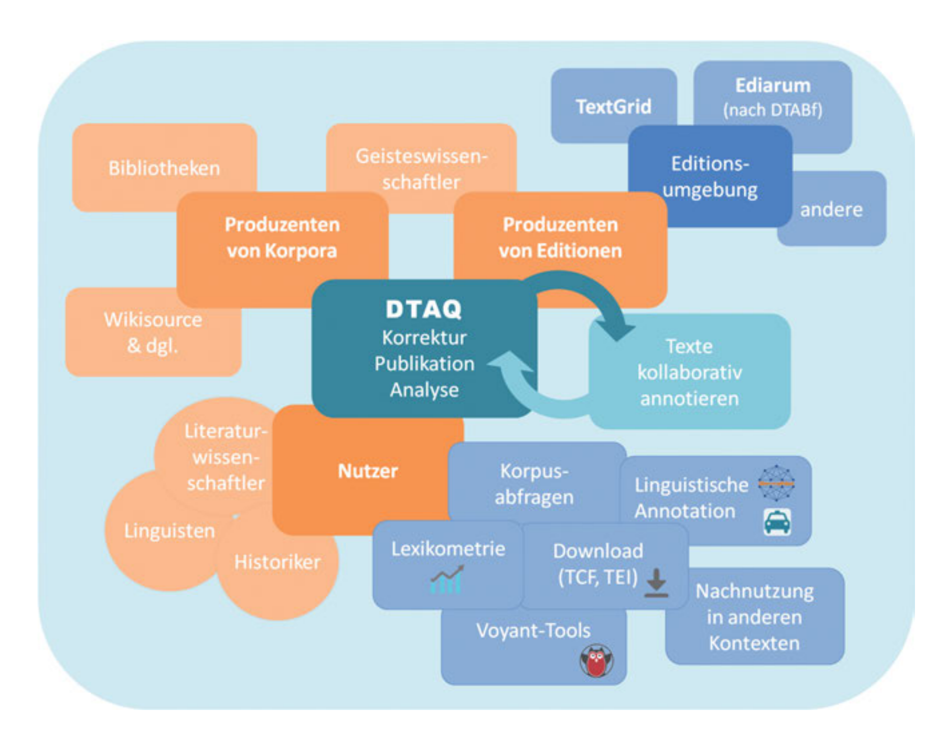


Abb. 10.1: Die Forschungsumgebung DTA.

Weiteres als Ganzes in wissenschaftlichen Kontexten nachgenutzt werden. Aufgrund der durch die Kodierung in DTABf gewährleisteten Interoperabilität können alle Texte des DTA darüber hinaus einfach in verschiedene Formate konvertiert werden (siehe Abschnitt 2.1). Seit 2014 ist das DTA fest in die CLARIN-Infrastruktur eingebunden. Die Möglichkeiten der Kooperation und der Nachnutzungen konnten dadurch weiter ausgebaut werden (siehe Abschnitt 4.2). Das DTA hat sich seitdem zu einer vollwertigen Forschungsplattform entwickelt, zu der Nutzerinnen und Nutzer entweder als Korpusproduzenten beitragen oder die sie als Plattform für die Textanalyse verwenden können. Abbildung 10.1 fasst die verschiedenen Komponenten zusammen, in deren Zentrum DTAQ als Korrektur-, Publikations- und Analyseplattform steht. Auf der einen Seite befinden sich die verschiedenen Korpusproduzenten (Geistes- und Sozialwissenschaftler und -wissenschaftlerinnen, Bibliotheken und außerakademische Initiativen wie beispielsweise Wikisource), auf der anderen Seite stehen Editionsumgebungen und Produzenten von Editionen. Die ‚klassische' Nutzung von DTAQ besteht in der kollaborativen Annotierung von Texten. Alle Texte des DTA können jederzeit korrigiert und annotiert werden, und die stets aktuelle Fassung kann aus der Plattform exportiert werden. Die vierte Kompo-

nente stellt schließlich die Analyse dar, mit den bereits erwähnten Werkzeugen CAB und GermaNet zur linguistischen Annotation, den verschiedenen Analysewerkzeugen und den Exportformaten zur flexiblen Nachnutzung in anderen Kontexten.

2 Ressourcenaufbau und Annotation

2.1 Datengrundlage und Annotationsformat: DTA-Basisformat (DTABf)

Um den Ansprüchen des DTA für eine möglichst vorlagengetreue Transkription historischer Quellen gerecht zu werden und gleichzeitig die Erfassung detailreicher Metadaten und umfangreicher Annotationen logischer und layoutbezogener Strukturen zu ermöglichen, wurde das DTA-Basisformat (DTABf) – ein auf den P5-Richtlinien der *Text Encoding Initiative* (TEI) basierendes XML-Format[2] – entworfen. Das DTABf stellt eine echte Teilmenge der von der TEI vorgegebenen Richtlinien zur Kodierung von Textdokumenten dar, d. h. das Tagset der TEI wurde hinsichtlich der verfügbaren Elemente und Attribute reduziert und hinsichtlich der Attribut-Werte spezifiziert (Haaf, Geyken & Wiegand 2014/2015; Geyken et al. 2012). Dadurch ist die volle Kompatibilität auch mit anderen TEI-basierten Projekten gewährleistet.

Das DTABf-Annotationsschema (RNG[3]) für historische Drucke (und weitere Dokumentklassen wie Zeitungen und Handschriften, vgl. Haaf & Schulz 2014; Haaf & Thomas 2016 [2017]) bildet zusammen mit einer umfangreichen Dokumentation und einem Schematron-Regelsatz die Grundlage für die XML-Auszeichnung aller Werke im DTA. Mithilfe von Konvertierungstools können aus DTABf-Dokumenten zahlreiche weitere Formate für die Weiterverarbeitung mit linguistischen Werkzeugen, für Suchmaschinenindizes, zur Präsentation der Texte (z. B. Lesefassungen für verschiedene Medien) und zum Export (z. B. in Zitationsumgebungen, Graphdatenbanken oder im CLARIN-Kontext für WebLicht[4]) automatisch erzeugt werden.

Mit dem DTABf als Leitfaden für die Auszeichnung der vielfältigsten Phänomene in historischen Textressourcen ist eine Brücke geschaffen worden, um

2 TEI P5 Guidelines: http://www.tei-c.org/Guidelines/P5/ (letzter Zugriff: 6. 11. 2017).
3 RELAX NG (RNG): http://relaxng.org/ (letzter Zugriff: 6. 11. 2017).
4 WebLicht: https://weblicht.sfs.uni-tuebingen.de/ (letzter Zugriff: 6. 11. 2017).

auch Editionen im DTA nutzbar zu machen. Das DTABf wird von der Deutschen Forschungsgemeinschaft nicht nur für die Textauszeichnung in historischen Sprachkorpora sondern auch als Basisformat für Editionen empfohlen.[5] Die Quelldateien des DTABf-Schemas und der Dokumentation stehen unter einer freien Lizenz zur Nachnutzung zur Verfügung.[6]

2.2 Die kollaborative Qualitätssicherungsumgebung DTAQ

Jedes Werk im DTA bildet eine Einheit aus drei Komponenten: Metadaten, Bilddigitalisate und Textdigitalisat. In der DTA-Infrastruktur werden diese Daten in entsprechenden Datenbanken und Speichersystemen vorgehalten. Diese werden dann auf der Präsentationsebene miteinander verknüpft. Um den verschiedenen Zugangsweisen, Sichten und Darstellungsformaten gerecht zu werden, sind dem Einspielen eines Werks in die DTA-Infrastruktur zunächst (automatisierte) Vorarbeiten vorangestellt:

1. Die Metadaten werden in eine SQL-basierte Datenbank überführt. Dadurch wird später eine gezielte und äußerst flexible Suche über diese Daten möglich. Da der Bestand des DTA auch von anderen Webdiensten regelmäßig abgefragt wird, werden dafür Metadaten in den Formaten Dublin Core,[7] CMDI[8] und EPICUR[9] erstellt und diese über Schnittstellen wie OAI-PMH,[10] BEACON[11] etc. zur Verfügung gestellt.
2. Die Bilddigitalisate werden für die Darstellung im Web in das JPEG-Format konvertiert. Dabei werden verschiedene Auflösungen einer jeden Dokumentseite erstellt: Für Vorschaubilder, die Einzeldarstellung in der Text-Bild-Ansicht und hochauflösend, um Details genauer betrachten zu können.

---

5 Vgl. die Handreichungen DFG 2015a und DFG 2015b.

6 Zugänglich unter https://github.com/deutschestextarchiv/dtabf (letzter Zugriff: 6. 11. 2107); vgl. auch Haaf (2017).

7 Dublin Core Metadata Initiative: http://dublincore.org/ (letzter Zugriff: 6. 11. 2017).

8 Component MetaData Infrastructure (CMDI): https://www.clarin.eu/content/component-metadata (letzter Zugriff: 6. 11. 2017).

9 Enhancement of Persistent Identifier Services: Comprehensive Method for Unequivocal Resource Identification (EPICUR): http://www.dnb.de/DE/Wir/Projekte/Archiv/epicur.html (letzter Zugriff: 6. 11. 2017).

10 Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH): https://www.openarchives.org/pmh/ (letzter Zugriff: 6. 11. 2017).

11 BEACON: https://de.wikipedia.org/wiki/Wikipedia:BEACON (letzter Zugriff: 6. 11. 2017).

3. Das Textdigitalisat wird in einzelne Textseiten (jeweils ein XML-Dokument) zerlegt, da bei größeren Werken mit mehreren hundert Seiten eine Extraktion der entsprechenden Daten zur Laufzeit (z. B. während des Abrufs einer bestimmten Seite in der Text-Bild-Ansicht) nicht performant möglich wäre. Des Weiteren wird eine Reintextfassung (d. h. die reine Transkription mit Zeilenumbrüchen und Absätzen, aber ohne jegliche Annotation) und eine Dokumentfassung für die Indizierung mit der Suchmaschine DDC erstellt. Da die Seitenzerlegung verlustfrei erfolgt, lässt sich aus den Einzelseiten jederzeit wieder das gesamte XML-Dokument erzeugen.

DTAQ ist eine Webanwendung, die weit mehr als nur die Präsentation einzelner Werke ermöglicht. Sie beinhaltet eine Nutzerverwaltung, die mithilfe von Rollen und Rechten mehrere Ebenen von Zugriffen und Annotationsoptionen für verschiedene Nutzergruppen anbietet. Nutzerinnen und Nutzer von DTAQ registrieren sich mit einem personalisierten Account auf der Plattform und können dabei verschiedene Expertisen (Expertise in Literatur- bzw. Sprachgeschichte, Fremdsprachenkenntnisse, Fachkenntnisse in der Transkription mathematischer Formeln u. a.) angeben. Das ermöglicht es, bei Zweifelsfällen oder schwierigen Textstellen gezielt andere Nutzerinnen und Nutzer mithilfe des Ticketsystems anzusprechen und so kollaborativ an den Dokumenten zu arbeiten. Zudem ist so eine Arbeit im Team problemlos möglich, indem bestimmte Arten von Fehlern gezielt einzelnen Nutzer und Nutzerinnen zugewiesen werden können. Die Personalisierung ermöglicht es auch, die eigenen Wünsche in Hinsicht auf die Darstellung in DTAQ für jeden Account zu speichern, unter anderem die optimale Text- und Bildbreite oder die präferierte Textansicht u. v. a.

Als ‚aktives Archiv' ermöglicht es das Deutsche Textarchiv mittels DTAQ, dass Nutzerinnen und Nutzer auch Bearbeitungen an den Textdigitalisaten selbst vornehmen können. Die Ablage der XML-Quellen in einem Versionierungssystem (git[12]) gewährleistet dabei zu jeder Zeit die Transparenz und Nachvollziehbarkeit der Genese eines digitalen Werks. Mithilfe zweier Online-Editoren (im sogenannten WYSIWYG-Modus[13] oder auch direkt im XML-Quelltext) können Nutzerinnen und Nutzer nachträglich Fehler korrigieren sowie zusätzliche Annotationen hinzufügen.

12 git: https://git-scm.com/ (letzter Zugriff: 6. 11. 2017).

13 WYSIWYG: What you see is what you get (man kann also direkt in der Präsentationsform der Texte ändern).

Technisch steht hinter DTAQ ein Webframework (Perl[14], Catalyst[15]), das mithilfe einer PostgreSQL-Datenbank[16] und XML-/XSLT[17]-Tools ein komfortables, effizientes Arbeiten mit dem Dokumentenbestand ermöglicht.

2.3 Erstellen und Kuratieren DTABf-konformer Textressourcen

Das DTA bietet seinen Nutzerinnen und Nutzern eine Vielzahl von Hilfsmitteln, um (historische) Textressourcen zur Integration in die DTA-Infrastruktur von Grund auf neu zu erarbeiten bzw. bestehende Daten zu kuratieren und damit DTABf-konform zu machen. Diese Hilfsmittel decken alle Bereiche der Erstellung bzw. Bearbeitung digitaler Volltexte ab: Verschiedene Werkzeuge bzw. Webservices unterstützen Forscherinnen und Forscher von der Metadatenerfassung über die Transkription und weitere Annotation der Volltextdaten sowie deren Konvertierung in eine lesefreundliche HTML-Ansicht. Das DTA unterstützt somit den gesamten Lebenszyklus digitaler Dokumente von der Erstellung über die Datenkonvertierung und -kuration (*Digital Curation*, vgl. dazu im Kontext des DTA/CLARINs Thomas & Wiegand 2015) bis hin zur Publikation und Analyse und dient damit als Textbearbeitungs- und Analyseplattform für (historische) Textressourcen. Anschließend an die Integration erfolgt die automatisierte computerlinguistische Erschließung der Daten, die somit unmittelbar im Kontext des DTA-Korpus genutzt werden können (siehe Abschnitt 3). Durch diese und weitere Hilfsmittel, zusätzlich unterstützt durch die mit zahlreichen Beispielen versehenen Richtlinien zur Transkription[18] sowie die umfangreiche, ebenfalls mit Beispielen aus den DTA-Korpora illustrierte Dokumentation zum DTABf,[19] werden Nutzerinnen und Nutzer somit in die Lage versetzt, Schritt für Schritt ihre Textressourcen standardkonform aufzubereiten.

Um Primärquellen in DTABf-konformer Weise von Grund auf neu zu erfassen, bietet es sich an, mit der über die Seiten des DTABf bereitgestellten XML-Vorlagendatei[20] zu beginnen und diese sukzessive um die Metadaten und

---

14 The Perl Programming Language: https://www.perl.org/ (letzter Zugriff: 6. 11. 2017).
15 Catalyst Web Framework: http://www.catalystframework.org/ (letzter Zugriff: 6. 11. 2017).
16 PostgreSQL: https://www.postgresql.org/ (letzter Zugriff: 6. 11. 2017).
17 Extensible Stylesheet Language Transformations (XSLT): https://www.w3.org/TR/xslt (letzter Zugriff: 6. 11. 2017).
18 Siehe http://www.deutschestextarchiv.de/doku/basisformat/transkription (letzter Zugriff: 6. 11. 2017).
19 Siehe http://www.deutschestextarchiv.de/doku/basisformat (letzter Zugriff: 6. 11. 2017).
20 Siehe http://www.deutschestextarchiv.de/files/vorlage_basisformat.xml (letzter Zugriff: 6. 11. 2017).

annotierte Transkription der Quelle zu ergänzen. Die Vorlagendatei ist TEI-basiert und bildet die vorgeschriebene Grundstruktur jedes TEI-Dokuments[21] ab. In der Datei bereits eingebunden sind bereits das DTABf-RNG-Schema, gegen das die Dokumente hinsichtlich der verwendeten Elemente und Attribut-Wert-Paare sowie deren korrekter Strukturierung und Schachtelung validiert werden, und ein entsprechender Schematron-Regelsatz, der weitere formale Festlegungen enthält.[22] Bei handschriftlichen Vorlagen sollte das spezifizierte RNG-Schema zum DTABf für Manuskripte (DTABf-M)[23] eingebunden werden. Als zusätzliche Arbeitserleichterung steht ein Webformular zur Erstellung DTABf-konformer TEI-Header zur Verfügung, über das die relevanten Angaben komfortabel erfasst werden können und aus dem heraus schließlich per Knopfdruck ein vollständiger, in sich valider TEI-Header erzeugt werden kann.[24]

Auch für die auf die Erstellung des XML-Dokuments folgenden Arbeitsschritte bietet das DTA hilfreiche Tools und Anwendungen an: Beispielsweise können die im DTA entwickelten XSLT-Stylesheets, mit deren Hilfe die XML-Datei in ein HTML-Format umgewandelt wird, genutzt werden, um eine lesefreundliche, an die spätere Darstellung auf der DTA-Seite angelehnte ‚Vorschauansicht' des DTABf-Dokuments im Browser zu generieren. Andere kleine Tools dienen beispielsweise dazu, die fortlaufende Nummerierung der aufeinanderfolgenden `<pb/>`-Elemente (d. h. die fortlaufende Nummerierung der Digitalisate der Vorlage) in der Datei (wieder-)herzustellen – eine ansonsten mühsame Handarbeit – oder die im Dokument enthaltenen Sonderzeichen[25] automatisch in die entsprechenden numerischen Entitäten für Zeichenverweise in der Form `&#xNNNN;` zu überführen.

21 Vgl. http://www.deutschestextarchiv.de/doku/basisformat/grundstrukturDokument (letzter Zugriff: 6. 11. 2017).

22 Vgl. allgemein Regular Language Description for XML New Generation (RELAX NG) und ISO Schematron, unter: http://relaxng.org/; http://schematron.com/ (letzter Zugriff: 6. 11. 2017). Speziell zu deren Verwendung und Handhabung innerhalb des DTABf die Dokumentation unter http://www.deutschestextarchiv.de/doku/basisformat/benutzungDTABfSchema (letzter Zugriff: 6. 11. 2017) sowie Haaf (2017).

23 Vgl. zum DTA-Basisformat für Manuskripte (DTABf-M) die Dokumentation unter http://www.deutschestextarchiv.de/doku/basisformat/manuskript (letzter Zugriff: 6. 11. 2017) sowie Haaf & Thomas (2016 [2017]).

24 Verfügbar unter http://www.deutschestextarchiv.de/dtae/submit/clarin (letzter Zugriff: 6. 11. 2017).

25 D. h. diejenigen, die außerhalb des Bereichs der am häufigsten verwendeten Unicode-Zeichen (kleinergleich U+00FF) liegen und daher oft zu Darstellungs-, Verarbeitungs- oder Dekodierungsfehlern führen.

In ähnlicher Weise wie für die im bisherigen Teil dieses Abschnitts beschriebene Neuerstellung DTABf-konformer Ressourcen können die vorgestellten Hilfsmittel – allen voran das RNG-Schema und der Schematron-Regelsatz, ggf. aber auch die hier vorgestellten Skripte, Tools und Webservices – zur Kuratierung externer Ressourcen genutzt werden. Dabei werden externe Textressourcen in das DTABf konvertiert und anschließend via DTAQ in die Korpusinfrastruktur integriert. Beispielsweise wurden im Rahmen eines von CLARIN-D geförderten Kurationsprojekts historische Textressourcen des 15.–19. Jahrhunderts aus verschiedenen Quellen in das DTA integriert.[26] (Konkrete Beispiele für integrierte externe Ressourcen finden sich in Abschnitt 4.2.)

Ein weiteres, im Rahmen von DTAQ bereitgestelltes Hilfsmittel dient dazu, die auf dem beschriebenen Weg erstellten, DTABf-konform annotierten Dokumente hinsichtlich der historischen Schreibweisen zu ‚normalisieren'.[27] Dieser Service spricht den CAB-Webservice[28] an, um für die historischen Schreibweisen im Text ein modernes Äquivalent zu ermitteln. Die Anreicherung des Textes mit den modernisierten Formen wird gemäß editorischen Konventionen mit Hilfe der TEI-Elemente `<choice>`, `<orig>` und `<reg>` dokumentiert. Das Element `<reg>` wird dabei zusätzlich mit dem Attribut-Wert-Paar `@resp="#cab"` als automatisierter, d. h. eben vom Webservice CAB verantworteter Eingriff gekennzeichnet. Beispielsweise wird die Transkription der Phrase „EJne feſte Burgk iſt vnſer GOtt/ ꝛc."[29] vollautomatisch annotiert als:

```
<choice><orig>EJne</orig>
    <reg resp="#cab">Eine</reg></choice>
<choice><orig>feſte</orig>
    <reg resp="#cab">feste</reg></choice>
<choice><orig>Burgk</orig>
    <reg resp="#cab">Burg</reg></choice>
<choice><orig>iſt</orig>
```

26 Vgl. zu diesem Projekt http://deutschestextarchiv.de/doku/clarin_kupro_publikationen (letzter Zugriff: 6. 11. 2017) bzw. Thomas & Wiegand (2015).

27 Dieser Webservice sowie alle im vorhergehenden Absatz erwähnten und weitere DTAQ-Tools sind verfügbar unter http://www.deutschestextarchiv.de/dtaq/tool (letzter Zugriff: 6. 11. 2017).

28 Vgl. DTA::CAB Web Service v1.82, http://www.deutschestextarchiv.de/cab/ (letzter Zugriff: 6. 11. 2017).

29 Vgl. zu diesem leicht modifizierten Textbeispiel [N. N.]: Jubilaeum Typographorum Lipsiensium Oder Zweyhundert-Jähriges Buchdrucker JubelFest. [Leipzig], 1640, S. [19]. In: Deutsches Textarchiv, http://www.deutschestextarchiv.de/oa_jubilaeum_1640/27 (letzter Zugriff: 6. 11. 2017).

```
    <reg resp="#cab">ist</reg></choice>
<choice><orig>vnſer</orig>
    <reg resp="#cab">unser</reg></choice>
<choice><orig>GOtt</orig>
    <reg resp="#cab">Gott</reg></choice>/
<choice><orig>ꝛc.</orig>
    <reg resp="#cab">etc.</reg></choice>
```

Auf diese Weise können auf denkbar einfache Art und unabhängig von bzw. noch vor der Integration der Texte in die DTA-Infrastruktur schreibweisennormierte Fassungen der vorlagengetreuen Transkriptionen historischer Textzeugen erstellt werden, die dann wiederum zur Bearbeitung mit externen, für gegenwartssprachliche Texte optimierten Anwendungen – beispielsweise zur Eigennamenerkennung oder Topic Modeling – genutzt werden können.

2.4 Arbeiten in DTAQ

Die Veröffentlichung der Ressourcen erfolgt zunächst auf der passwortgeschützten DTAQ-Plattform. Hier findet die summative Qualitätssicherung[30] statt, d. h. die Text-Bild-Zuordnung, die Transkription und Annotation usw. können webbasiert und kollaborativ geprüft werden. Für jedes Dokument wird eine eigene Startseite erstellt, auf der die bibliographischen Metadaten und weitere relevante Informationen zur Quelle, einschließlich der zugrunde gelegten Text- und Bildvorlagen, der Lizenz für die Nachnutzung, der Korpuszugehörigkeit des Dokuments, der Bearbeiter der digitalen Edition sowie die dabei ggf. verbliebenen Abweichungen der Transkription bzw. der Annotation von den DTA-Vorgaben zusammengefasst sind. Zu beiden Bereichen, d. h. den allgemeinen Informationen und den bibliographischen Metadaten, können über einen Browserdialog Anmerkungen und Berichtigungen hinzugefügt werden. Unter *Ansichten* stehen in der Korrekturumgebung DTAQ bzw., sobald das Dokument freigeschaltet ist, auf der DTA-Webseite neben dem Zugang zur Text-Bild-Ansicht unter anderem verschiedene, automatisch aus dem TEI-XML generierte Download-Formate sowie eine Reihe analytischer Zugänge bereit. Mit der linguistischen Suchmaschine DDC und der grep-Suche werden parallel zwei Volltextsuchen angeboten.

---

30 Vgl. für eine ausführliche Darstellung der formativen und summativen Qualitätssicherung im DTA Geyken et al. 2012.

Ein automatisch aus der <div>-Strukturierung des Dokuments erzeugtes *Inhaltsverzeichnis* erlaubt das gezielte Ansteuern bestimmter Textpassagen. Durch Klick auf einen Inhaltsabschnitt, auf die Ansicht *Korrekturumgebung* oder auf das Faksimile des Titelblatts gelangt man von dieser Startseite aus zur Text-Bild-Ansicht in DTAQ. Neben dem Faksimile der Vorlage in der ersten Spalte der dreiteiligen Ansicht wird in der mittleren Spalte die Transkription des Textes der entsprechenden Seite angezeigt. Für die Textansicht sind verschiedene Optionen wählbar: die XML-Ansicht, eine reine Textansicht, die CAB-Ansicht mit der automatisch durchgeführten orthographischen Normalisierung der zugrundeliegenden Transkription, eine erweiterte Ansicht der Part-of-Speech-Analyse sowie der Lemmatisierung.

Zur Überprüfung des Texts eignet sich die standardmäßig angezeigte HTML-Fassung am besten.[31] In dieser lesefreundlichen, per XSLT aus dem XML erstellten Textansicht lassen sich nun die bei dem jeweiligen Korrekturgang gefundenen Transkriptions-, Auszeichnungs- oder Druckfehler mittels sogenannter Tickets melden. Die gemeldeten Tickets enthalten neben dem Fehlertyp, der fehlerhaften Textstelle, einem Vorschlag zur Behebung des Fehlers und ggf. weiteren Kommentaren zugleich auch die Angaben zum Zeitpunkt der Erfassung und zum DTAQ-Account, durch den die Meldung angelegt wurde. Beobachtungen, die sich nicht nur auf die jeweilige Seite, sondern das gesamte Dokument beziehen, können entsprechend gekennzeichnet werden. Darüber hinaus ist es möglich, dem Ticket eine mehr oder weniger hohe Priorität zuzuordnen und dieses einer bestimmten Bearbeiterin bzw. einem bestimmten Bearbeiter zuzuweisen. Anschließend wird der Korrekturstatus der Seite gekennzeichnet, d. h. ob die Überprüfung auf der Ebene des Textes, eines punktuellen Vergleichs von Transkription und Faksimile und/oder auf der Ebene der XML-Annotation durchgeführt wurde. Die gemeldeten Fehler werden so anhand der Beschreibungen in den Tickets vom DTA-Team geprüft und entsprechend direkt in den XML-Dokumenten behoben.

Eine andere Möglichkeit, die auch durch DTA-externe Nutzerinnen und Nutzer zur sukzessiven Korrektur bzw. Kuratierung der Ressourcen direkt in der webbasierten DTAQ-Oberfläche genutzt werden kann, sind die beiden integrierten Online-Editoren, die optional bereitgestellt werden. Der einfache WYSIWYG-Editor bietet die Möglichkeit, innerhalb der HTML-Ansicht Änderungen auf der Textoberfläche, insbesondere also bei Transkriptionsfehlern, vorzunehmen. Die vorgenommenen Änderungen werden im Zuge der Speicherung automatisch dokumentiert und versioniert. Der parallel dazu angebotene XML-

31 Vgl. zu diesem Abschnitt abermals Geyken et al. (2012) sowie Haaf & Thomas (2016: insbes. 227 f.).

Editor erlaubt Änderungen, die (auch) das strukturelle und typographische Markup der Dokumente betreffen; hier können beispielsweise manuelle Normalisierungen vorgenommen oder Druckfehler in dokumentierter Weise mittels der TEI-Elemente <choice>, <sic> und <corr> behoben werden:



Abb. 10.2: XML-Editor in DTAQ mit geöffnetem Browserdialog (Wrap-Tag-Funktion), http://www.deutschestextarchiv.de/dtaq/book/view/humboldt_manati_1838?p=9&view=xmleditor (letzter Zugriff: 6. 11. 2017)

Neben der farblich differenzierten Syntaxhervorhebung und der Wrap-Tag-Funktion, die Nutzerinnen und Nutzer von etablierten Programmen wie dem oXygen XML Editor gewohnt sind, bietet die rechte Spalte eine Anzahl häufig genutzter Sonderzeichen und Tags zur Auswahl, die die Bearbeitung der Dokumente direkt im XML-Modus erleichtern. Vorgenommene Änderungen auf der XML-Ebene werden zunächst auf ihre Wohlgeformtheit und auf ihre Validität hin gegen das DTABf-Schema geprüft, um sicherzustellen, dass keine in dieser Hinsicht ungültigen Änderungen in den XML-Dokumenten vorgenommen werden; anschließend werden die formal korrekten Änderungen im XML-Dokument gespeichert und ebenfalls versioniert.

Durch die beiden webbasierten, direkt in DTAQ eingebundenen Editoren lassen sich nun nicht nur die bereitgestellten Quellen sehr viel komfortabler und mit Unterstützung auch DTA-externer Nutzerinnen und Nutzer korrigieren. Sie ermöglichen darüber hinaus auch eine effiziente und fortlaufende Verfeinerung bzw. Vertiefung der Annotation insgesamt. So wurden beispielsweise im

Kooperationsprojekt *Hidden Kosmos*[32] zunächst alle Manuskripte des Korpus DTABf-konform transkribiert, hinsichtlich der wesentlichen Struktur- und graphematischen Merkmale ausgezeichnet und in DTAQ publiziert. Anschließend erfolgte nicht nur die Qualitätssicherung und Korrektur von Transkriptions- und Auszeichnungsfehlern direkt in DTAQ, sondern wurden innerhalb des etwa 3.500 Seiten umfassenden Korpus auch sämtliche vorkommenden Personennamen explizit ausgezeichnet. Die insgesamt mehr als 8.000 Personennamen wurden mit einem `<persName>`-Tag versehen und, soweit möglich, mittels des `@ref`-Attributs mit einem eineindeutigen Identifizierer aus einem Normdatensatz[33] versehen. Das so erzeugte, umfassende Personenregister[34] konnte dank des webbasierten XML-Editors ortsunabhängig vom Projektteam an der Humboldt-Universität erstellt werden. In ähnlicher Weise wurden innerhalb des DFG-Projekts *AEDit Frühe Neuzeit*[35] etwa 340 Leichenpredigten mit mehr als 16.000 Seiten im Team der Kooperationspartner BBAW, Herzog August Bibliothek Wolfenbüttel (HAB) und der Forschungsstelle für Personalschriften an der Philipps-Universität Marburg von den verschiedenen Standorten aus kollaborativ online bearbeitet.

3 Analysewerkzeuge im DTA

3.1 Linguistische Datenanalyse im DTA

Für bestimmte Anwendungsbereiche der Auswertung der Korpora im DTA werden bereits über die DTA-Plattform eigene Werkzeuge angeboten. Auf diese Weise können die DTA-Daten hinsichtlich verschiedener Phänomentypen analysiert werden, ohne dass sie heruntergeladen und mit externen Tools ver-

---

32 Vgl. Humboldt-Universität zu Berlin: *Hidden Kosmos – Reconstructing Alexander von Humboldt's "Kosmos-Lectures"*, http://www.culture.hu-berlin.de/hidden-kosmos (letzter Zugriff: 6. 11. 2017).

33 Genutzt wurde vorzugsweise die Gemeinsame Normdatei (GND), vgl. http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html (letzter Zugriff: 6. 11. 2017); sofern keine Normdaten zu den betreffenden Personen in der GND verzeichnet waren, wurden alternativ die Datensätze aus dem Virtual International Authority File, https://viaf.org/ (letzter Zugriff: 6. 11. 2017) oder Wikidata, https://www.wikidata.org/wiki/Wikidata:Hauptseite (letzter Zugriff: 6. 11. 2017), verwendet.

34 Verfügbar unter http://www.deutschestextarchiv.de/kosmos/person (letzter Zugriff: 6. 11. 2017).

35 Vgl. dazu http://www.deutschestextarchiv.de/doku/textquellen#aedit (letzter Zugriff: 6. 11. 2017).

knüpft werden müssen. Das Potential der linguistischen Analyse auf der DTA-Plattform wird im Folgenden umrissen.

Sämtliche Texte des DTA durchlaufen vollautomatisch eine Reihe linguistischer Verarbeitungsschritte, welche durch die Software CAB geleistet werden. CAB umfasst die Satzsegmentierung, Tokenisierung, Lemmatisierung, Modernisierung historischer Schreibweisen sowie das Part-of-Speech-Tagging gemäß STTS (Jurish 2012; Jurish & Würzner 2013).

Die Ergebnisse der linguistischen Analyse können mithilfe der Suchanfragesprache DDC[36] in die Korpusanalyse eingebunden werden. DDC ermöglicht die Formulierung komplexer Suchanfragen (Jurish, Thomas & Wiegand 2014).[37] Insbesondere können damit linguistische Annotationen auf Wortposition (Lemmatisierung, POS-Tagging) mit Phrasensuche, Boole'schen Operatoren und regulären Ausdrücken verknüpft werden.

Um die Analyseergebnisse für ein Token einzusehen sowie um zu überprüfen, welches Lemma, Wortart bzw. welche orthographischen Varianten für dieses Token automatisch ermittelt wurden, kann die Benutzeroberfläche des CAB-Webservice[38] konsultiert werden. Hier ist z. B. einsehbar, dass das Token „Frewde“ dem Lemma „Freude“ sowie dem POS-Tag „NN“ (Appellativum) zugeordnet wird und es auf eine Vielzahl von möglichen orthographischen Varianten abgebildet wird, von „Frewdt“ über „fräud“ und „Freüd“ bis hin zu „fröude“, „fröide“ oder „vröide“.[39] Vor allem aber kann der CAB-Webservice genutzt werden, um eigene Texte mit der CAB-Software zu analysieren. Für einen schnellen Einblick in die orthographische Normierung und das POS-Tagging im größeren Textzusammenhang steht die CAB- bzw. POS-Ansicht zu jeder Buchseite in DTAQ zur Verfügung.

---

36 Dialing DWDS Concordancer (DDC), http://www.deutschestextarchiv.de/doku/software#ddc; vgl. Jurish, Thomas & Wiegand 2014. Zu den vielfältigen Möglichkeiten der Volltextsuche im DTA mit der DDC-Suchmaschine siehe auch die Hilfeseite http://www.deutschestextarchiv.de/doku/DDC-suche_hilfe (letzter Zugriff: 6. 11. 2017).

37 Für die DDC-basierte Recherche in allen Korpora des Deutschen Textarchivs steht die Suchmaske unter http://kaskade.dwds.de/dstar/dta (letzter Zugriff: 6. 11. 2017) zur Verfügung. Hier lässt sich einschränken, ob im Gesamtkorpus oder in ausgewählten Teilkorpora recherchiert werden soll (sog. Flags). Die Korpora umfassen auch alle Texte in DTAQ. Zur DDC-Syntax vgl. als Einstieg http://www.deutschestextarchiv.de/doku/DDC-suche_hilfe (letzter Zugriff: 6. 11. 2017) sowie weiterführend http://odo.dwds.de/~jurish/software/ddc/querydoc.html (letzter Zugriff: 6. 11. 2017).

38 Vgl. Anm. 27; Einstieg und Links zur Dokumentation: http://odo.dwds.de/~moocow/software/DTA-CAB/doc/html/DTA.CAB.WebServiceHowto.html (letzter Zugriff: 6. 11. 2017).

39 Vgl. http://www.deutschestextarchiv.de/demo/cab/query?a=expand&fmt=text&raw=1&q=Frewde (letzter Zugriff: 6. 11. 2017).

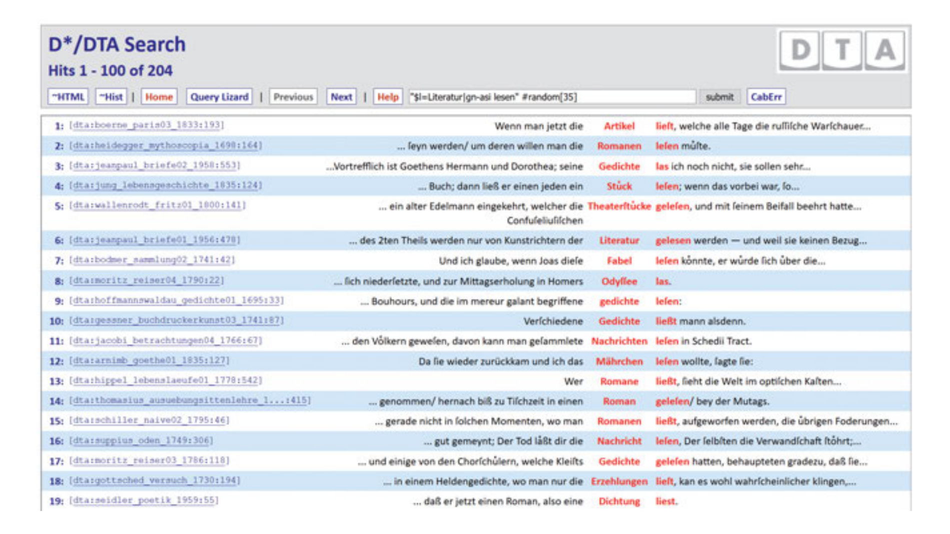


Abb. 10.3: DDC-Abfrage: `"$l=Literatur|gn-asi lesen"`.

Alle Texte der DTA-Korpora wurden mit einer Thesaurus-Funktion basierend auf dem Wortnetz GermaNet versehen. Über die Verknüpfung der Lemmata mit SynSets ist daher nun eine semantische Recherche möglich. Konkret können für ein Lemma dessen Hyperonyme, Hyponyme oder Synonyme im Korpus recherchiert werden (Abb. 10.3).

Valide Expansionen sind dabei `"gn-asi"` (sämtliche Hyponyme zum gegebenen Lemma), `"gn-isa"` (entsprechend die Hyperonyme), `"gn-syn"` (entsprechend die Synonyme). Die Suche kann auch auf die Hypero- und Hyponyme bestimmter Ordnungen (`"gn-asi1"`, `"gn-asi2"`) eingeschränkt werden. Neben GermaNet ist darüber hinaus auch der OpenThesaurus[40] an die DTA-Korpora angebunden, sodass äquivalente Recherchen auf dem Datenmaterial dieses Thesaurus möglich sind.[41]

3.2 Visualisierung des zeitlichen Verlaufs mit Verlaufskurven

Über die reine Korpusrecherche hinaus sind verschiedene lexikometrische Analysen über die DTA-Plattform möglich. So lässt sich etwa die relative Verteilung von (komplexen) Ausdrücken in den DTA-Korpora mit Hilfe einer Wort-

40 Vgl. https://www.openthesaurus.de (letzter Zugriff: 6. 11. 2017).
41 Für eine Dokumentation zu den Thesaurus-Funktionalitäten im DTA vgl. http://odo.dwds.de/~moocow/software/ddc/querydoc.html#dta_expand_gn (letzter Zugriff: 6. 11. 2017) bzw. http://odo.dwds.de/~moocow/software/ddc/querydoc.html#dta_expand_ot (letzter Zugriff: 6. 11. 2017).

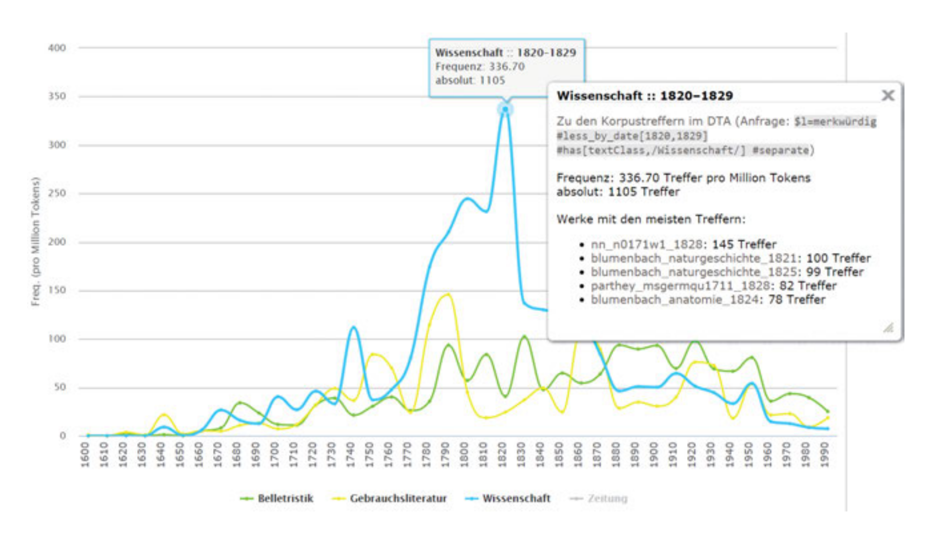


Abb. 10.4: Wortverlaufskurve für DDC-Abfrage: `$l=merkwürdig`.

verlaufskurve darstellen, welche auf normierten Frequenzberechnungen und Glättungsverfahren beruht (Geyken et al. 2015).[42] Für das Lemma „merkwürdig" wird so z. B. auf einen Blick deutlich, dass es in der Wissenschaftssprache des 19. Jahrhunderts signifikant häufig verwendet wurde (Abb. 10.4). Neben der Standardansicht kann auch eine Ansicht mit den Rohfrequenzen gewählt werden. In dieser lassen sich für jedes Jahr und jede Textsorte die absoluten Frequenzen nachvollziehen. Schließlich ermöglicht die erweiterte Ansicht die genaue Einstellung aller Parameter, wie beispielsweise der Zeitintervalle, der Glättungsumgebung oder des Konfidenzintervalls zur Bestimmung von Ausreißern. Eine weitere Besonderheit der Wortverlaufskurve besteht darin, dass alle Ergebnisse an die Korpora zurückgebunden werden. Durch Klick auf jeden Messpunkt gelangt man zu den Korpus-Konkordanzen. Im Falle der Abbildung 10.4 kann man somit beispielsweise durch Klick auf das Maximum im Zeitintervall 1820–1829 erfahren, dass unter anderem die Nachschriften der Kosmos-Vorlesungen Alexander von Humboldts sowie Werke Johann Friedrich Blumenbachs hinter diesen hohen Trefferzahlen stehen.

---

42 Die Wortverlaufskurve ist zugänglich unter http://www.deutschestextarchiv.de/search/plot/ (letzter Zugriff: 6. 11. 2017).

3.3 Kollokationsanalyse mit DiaCollo

Die weiterführende Untersuchung der Entwicklung von Wörtern in ihren lexikalischen Kontexten im zeitlichen Verlauf ist dann mit dem Werkzeug DiaCollo möglich (Jurish 2015; Jurish, Geyken & Werneke 2016, Lemnitzer, Jurish & Burkhardt 2016). DiaCollo ermittelt die statistisch signifikanten Kollokationen zu einem gegebenen Ausdruck in verschiedenen Zeitschnitten und visualisiert ihre Signifikanz mit einem Farbschema. Die Analyse kann dabei individuell parametrisiert werden. Zum Beispiel ist es möglich, die Größe der Zeitschnitte anzupassen, die Kollokate auf bestimmte POS-Tags zu beschränken oder die Menge der einbezogenen stärksten Kollokate (k-best-Wert) zu variieren. Verschiedene Visualisierungen stehen zur Verfügung, wie beispielsweise eine *bubble*-Ansicht, eine Ansicht als Schlagwortwolke oder die von Google entwickelte *gmotion*-Ansicht.

Nutzt man z. B. DiaCollo, um die Nomen zu ermitteln, welche typische Kollokationen des Tokens „merkwürdig“ sind, so zeigt sich zunächst, dass ganz generell die Menge solcher typischen Nomen-Kollokationen zunimmt (d. h. dass sich die Kontexte, in welchen dieses Adjektiv typischerweise verwendet werden konnte, vervielfältigen). Des Weiteren wird deutlich, dass zwischen 1750 und 1790 die „merkwürdige Begebenheit“ eine recht typische Wortverbindung war, während ab den 1820er Jahren bemerkenswert häufig von der „merkwürdigen Erscheinung“ die Rede ist (Abb. 10.5). Ebenso wie bei den in Ab-

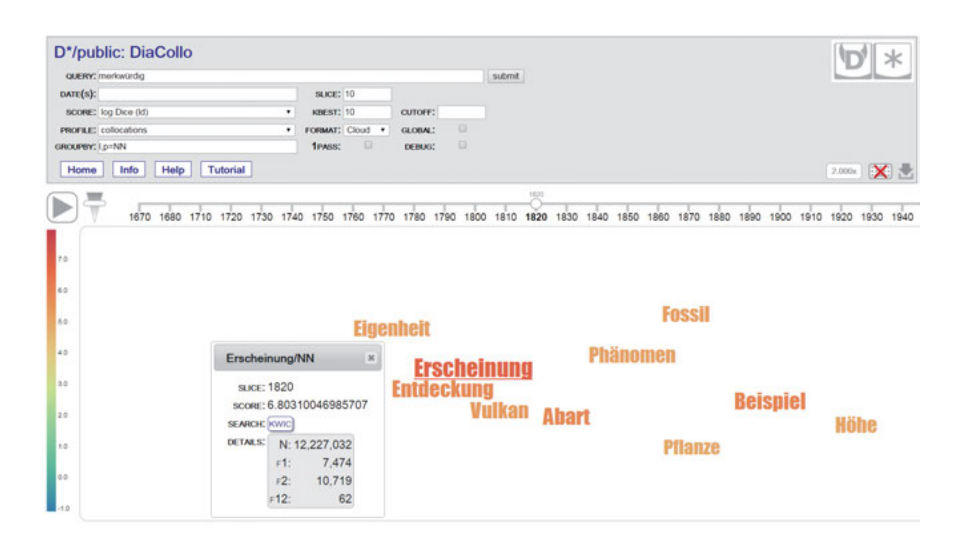


Abb. 10.5: DiaCollo: typische Nomen-Kollokationen (GROUPBY: l,p = NN) zu „merkwürdig“ im Zeitschnitt 1820.

schnitt 3.2 beschriebenen Wortverlaufskurven ermöglicht auch DiaCollo die Rückbindung an die einzelnen Korpuskonkordanzen durch Klick auf die Kollokate. In Abbildung 10.5 würde beispielsweise der Klick auf das Kollokat „Erscheinung" zu einer Trefferliste mit den Konkordanzen führen.

3.4 Quantitative Einzeltextanalyse mit Voyant Tools

Schließlich werden die einzelnen Dokumente an die Voyant Tools[43] angebunden, die für das jeweilige DTA-Werk über einen entsprechenden Unterpunkt auf der Buchstartseite erreicht werden können. Die XML-Volltexte aus dem DTA werden eigens zu diesem Zweck und ohne weiteren nutzerseitigen Aufwand präprozessiert, um eine nahtlose Verwendung und optimale Analyseergebnisse zu gewährleisten. Zur Analyse mit Voyant stellt das DTA drei spezielle XML-Fassungen zur Verfügung:

1. Eine *zeichennormierte Fassung* (unicruftxml): Diese XML-Fassung bietet den Text in transliterierter Orthographie, d. h. in einer Fassung, in der alle Zeichen, die außerhalb der Latin-1-Kodierung (ISO/IEC 8859-1) liegen, durch Zeichen innerhalb von Latin-1 approximiert werden. Damit sind Probleme bei der Voyant-seitigen Behandlung von Zeichen wie dem ‚langen s' (ſ, U+017F) oder dem ‚hochgestellten e' (U+0364) zur Kennzeichnung von Umlauten ausgeschlossen. Abgesehen davon bleiben die Graphie der Vorlage und auch die Silbentrennung am Seiten- und Zeilenende erhalten.
2. Eine *hinsichtlich der Schreibweisen normierte Fassung* (normxml): Diese XML-Fassung bietet den Text ebenfalls Latin-1-approximiert (siehe 1.) und zusätzlich in normalisierter Orthographie basierend auf der Verarbeitung mit CAB. In diesem Zuge wird auch die Silbentrennung am Seiten- und Zeilenumbruch aufgelöst.
3. Eine *lemmatisierte Fassung* (lemmaxml): Diese XML-Fassung bietet den Text in lemmatisierter Forma, wobei für die Lemmata ebenfalls mit normierten Zeichen (nach 1.) und modernisierter Orthographie (nach 2.) wiedergegeben werden.

Über die Voyant Tools sind nun verschiedene Frequenzanalysen auf Dokumentebene visualisierbar, so etwa Term- und Phrasenfrequenzen sowie Frequenzentwicklungen für Terme (Trends) im Verlauf des Dokuments (Sinclair & Rockwell 2017). So können Terme hinsichtlich ihrer Bedeutung für das jeweilige Werk analysiert werden (Vgl. z. B. Bird, Menzies & Zimmermann 2015: 51).

43 Vgl. https://voyant-tools.org/ (letzter Zugriff: 14. 5. 2018).



Abb. 10.6: Voyant-Analyse von Keller, Gottfried: Der grüne Heinrich. Bd. 1. Braunschweig, 1854, basierend auf der normalisierten Textfassung, unter: http://www.deutschestextarchiv.de/book/download_normxml/keller_heinrich01_1854 (letzter Zugriff: 6. 11. 2017).

4 Interdisziplinäre Vernetzung des DTA

4.1 Verwertung von DTA-Daten in externen wissenschaftlichen Kontexten

Die vorgestellten Analysewerkzeuge der DTA-Plattform ermöglichen bereits die Auswertung der DTA-Korpora in vielerlei Hinsicht. Zudem ist die weitere Nachnutzung der Daten mithilfe externer Werkzeuge, die z. B. über die CLARIN-Infrastruktur bereitgestellt werden, möglich und selbstverständlich auch erwünscht. Wesentliche Voraussetzung dafür ist zum einen die Interoperabilität der Daten untereinander, sodass diese vollautomatisch und jederzeit mit geringem Aufwand in die jeweiligen Eingabeformate konvertiert werden können. Zum anderen ist diese Form der Nachnutzung besonders gut möglich, wenn sowohl bei den Daten als auch bei den Eingabeformaten der Werkzeuge standardisierte Formate zur Anwendung kommen, die dokumentiert und leicht zugänglich sind. Diese beiden Voraussetzungen werden für die Korpusdaten des DTA durch das DTA-Basisformat erfüllt, das eine einheitliche und eindeutige Richtlinie für die Textauszeichnung darstellt und dabei auf den weit verbreiteten TEI-Richtlinien basiert.

Für die freie Nachnutzung in externen Kontexten werden die DTA-Texte einzeln und als Korpus zum Download zur Verfügung gestellt.[44] Da sämtliche Daten einheitlich entsprechend dem DTA-Basisformat kodiert sind, können sie mit geringem Aufwand in andere Formate konvertiert werden. Über die DTA-Plattform werden die Texte außerdem bereits im TCF (*Text Corpus Format*; Heid et al. 2010) bereitgestellt, dem Eingangsformat für die CLARIN-Services (Eckart 2012). Die bereitgestellten TCF-Dateien können zum Beispiel in die WebLicht-Plattform (Hinrichs, Hinrichs & Zastrow 2010)[45] hineingeladen und dort mit den Tools verschiedener CLARIN-Zentren weiter analysiert werden. So stehen über WebLicht zum Beispiel weitere Tools zur linguistischen Vorverarbeitung (Tokenisierung, Lemmatisierung, POS-Tagging etc.), verschiedene Syntaxparser (z. B. Stanford Parser) sowie Tools zur morphologischen Analyse und zur Eigennamenerkennung zur Verfügung. Bei der Nutzung von WebLicht können verschiedene Werkzeuge in Folge zur Anwendung kommen, wobei die Zusammensetzung der jeweiligen Werkzeugkette individuell bestimmt werden kann. Um genuine TEI-XML-Dateien in WebLicht weiterverarbeiten zu können, stellt die BBAW einen Converter als Webservice über WebLicht bereit, der die Konvertierung aus TEI nach TCF und die Rückkonvertierung nach TEI im Anschluss an die WebLicht-Analyse leistet.[46]

Des Weiteren sind die DTA-Korpora an die Föderierte Volltextsuche (*Federated Content Search*)[47] von CLARIN angebunden und somit gemeinsam mit anderen CLARIN-Korpora direkt recherchierbar. Die Verbindung der DTA-Korpusdaten mit anderen Korpora im CLARIN-Verbund wird außerdem über die CLARIN-übergreifende Metadatenplattform VLO (*Virtual Language Observatory*)[48] erreicht, welche durch die Bereitstellung von CMDI-Metadatensätzen zum Harvesting bedient wird (Wittenburg & Uytvanck 2012).

Auch in anderen Umgebungen außerhalb CLARINs ist die Nachnutzung der DTA-Daten möglich und erprobt. Metadaten werden in den genannten Formaten und zusätzlich im Dublin-Core-Format über die OAI-PMH-Schnittstelle des DTA bereitgestellt und beispielsweise von der Europeana[49] und der Biele-

44 Einzelne Werke können über die jeweilige Buchstartseite in verschiedenen Formaten heruntergeladen werden; der Download des Korpus ist unter http://www.deutschestextarchiv.de/download (letzter Zugriff: 6. 11. 2017) möglich.

45 Zugänglich unter: https://weblicht.sfs.uni-tuebingen.de/ (letzter Zugriff: 6. 11. 2017).

46 Außerhalb WebLicht zugänglich unter: http://kaskade.dwds.de/tei-tcf/ (letzter Zugriff: 6. 11. 2017).

47 Zugänglich unter: https://www.clarin.eu/contentsearch (letzter Zugriff: 6. 11. 2017).

48 Zugänglich unter: http://vlo.clarin.eu (letzter Zugriff: 6. 11. 2017).

49 Vgl. http://www.europeana.eu (letzter Zugriff: 6. 11. 2017).

feld Academic Search Engine (BASE)[50] abgefragt. Darüber hinaus harvesten einige Bibliotheken, welche die physischen Quellen der Werke im DTA vorhalten, Informationen zu den entsprechenden Werken im DTA und verlinken diese über ihren OPAC.[51]

Die DTA-Korpusdaten wurden außerdem bereits in andere Korpustools eingespeist. So wurde die TCF-Fassung des DTA-Korpus für ein entsprechendes Korpus-Add-on des Werkzeugs *Corpus Explorer* nachgenutzt, wobei die Informationen zu Tokens, Lemmata, POS und Orthographie ausgewertet wurden.[52] Auch der *Corpus Explorer* bietet darauf aufbauend Syntax Parsing, Kookkurenzanalysen, Verlaufskurven u. a. an. Die Möglichkeit, mit CAB analysierte Daten in das Korpuswerkzeug TXM zu integrieren und dort weiter zu verarbeiten, wurde für Karl Philipp Moritz' Werk *Anton Reiser* prototypisch realisiert.[53] Eine aktuelle Entwicklung bildet die Weiterentwicklung des Tools JCore, eines Werkzeugs zur Nutzung von NLP-Prozessketten in UIMA[54], ursprünglich mit Fokus auf englischsprachiger biomedizinischer Wissenschaftsliteratur, für die deutsche Sprache. In diesem Zusammenhang entstand der *DTA Collection Reader*[55] (Hahn et al. 2016; Hellrich, Matthies & Hahn 2017).

Das DTA-Kernkorpus wurde zudem in das Leipziger Werkzeug CTS (*Canonical Text Service Protocol*) integriert, wo es mit feingranularen persistenten IDs für die Zitation versehen wird und mit darauf aufbauenden Software-Werkzeugen (z. B. zur Alignierung von Textstellen unterschiedlicher Korpora) weiterverarbeitet werden kann (Tiepmar et al. 2016; Tiepmar et al. 2017). Ferner wurde die TCF-Fassung des DTA-Korpus experimentell in eine Graph-Datenbank integriert (Kuczera 2017). Weiterhin bilden 633 Werke der

50 Vgl. https://www.base-search.net/ (letzter Zugriff: 6. 11. 2017).

51 So etwa die Staatsbibliothek zu Berlin (SBB-PK, http://staatsbibliothek-berlin.de/ [letzter Zugriff: 6. 11. 2017]) oder die Staats- und Universitätsbibliothek Göttingen (https://www.sub.uni-goettingen.de [letzter Zugriff: 6. 11. 2017]); vgl. z. B. die entsprechende Verlinkung Kants *Critik der reinen Vernunft* (1781, http://www.deutschestextarchiv.de/kant_rvernunft_1781), abgerufen am 31. 3. 2017 im OPAC der SBB-PK: http://stabikat.sbb.spk-berlin.de/DB=1/XMLPRS=N/PPN?PPN=83453522X.

52 Vgl. http://notes.jan-oliver-ruediger.de/korpora/; http://notes.jan-oliver-ruediger.de/dta-kernkorpus-als-korpus-addon-verfuegbar/ (letzter Zugriff: 6. 11. 2017).

53 Vgl. https://groupes.renater.fr/wiki/txm-users/public/umr_ihrim_moritz (letzter Zugriff: 6. 11. 2017).

54 Unstructured Information Management Architecture; http://uima.apache.org/ (letzter Zugriff: 6. 11. 2017).

55 Zugänglich unter: https://github.com/JULIELab/jcore-base/tree/master/jcore-dta-reader (letzter Zugriff: 6. 11. 2017).

literarischen Moderne aus den DTA-Korpora einen Teil des Korpus KOLIMO (Herrmann & Lauer 2017).[56]

Für die Editionswissenschaften wurde seitens der Arbeitsgruppe TELOTA der BBAW die Editionsumgebung ediarum[57] für das DTA-Basisformat angepasst. Mit ediarum können TEI-XML-basierte Editionen im Autormodus des oXygen-XML-Editors erarbeitet werden. Die Kodierung und Verwaltung der Editionsdaten erfolgt in einer eXist-Datenbank (Dumont & Fechner 2014/15). Im Rahmen des Vorhabens *Alexander von Humboldt auf Reisen* wird mit dieser Infrastruktur eine Edition nach DTA-Basisformat erstellt (Dumont et al. 2016).[58]

4.2 Nachnutzung externer wissenschaftlicher Daten im DTA

Durch die Etablierung digitaler Arbeitsmethoden steigt in den Geisteswissenschaften der Bedarf an digitalisierten Quellentexten. Dabei ist einerseits eine kritische Menge für wissenschaftlich fundierte Aussagen unerlässlich, während andererseits der Bedarf in der Datenmenge nicht zulasten der Qualität gehen darf. Was also benötigt wird, sind umfangreiche, hochwertige und standardisiert aufbereitete Textsammlungen und Korpora. Solche Daten können nur mit hohem Aufwand erzielt werden, was wiederum der Nachfrage entgegenzustehen scheint.

Auf der anderen Seite entstehen in unterschiedlichen Kontexten immer mehr solcher hochwertigen digitalen Daten, etwa im Rahmen von Editionen, in Form projekteigener oder für individuelle Forschungsarbeiten erarbeiteter Spezialkorpora oder als Textkollektionen der interessierten Community (z. B. Wikisource[59]). Gerade seitens der quantitativen Linguistik entstehen schon seit langem solcherlei digitale Forschungsdaten, die jedoch teilweise in veralteten und/oder stark individualisierten Formaten vorliegen. Die Herausforderung besteht nun darin, diese Forschungsdaten aus den verschiedenen Quellen zu ermitteln, zu standardisieren und nach einheitlichen Richtlinien aufzubereiten, sodass sie als einheitliches Gesamtkorpus auswertbar werden. Im Rahmen des Moduls DTAE verfolgt das DTA diese Aufgabe (Thomas & Wiegand 2015).[60] Dabei geht es um zweierlei: Einerseits werden Verfahren entwickelt, um Daten

---

56 Zugang zur KOLIMO-Plattform unter: https://kolimo.uni-goettingen.de/index.html (letzter Zugriff: 6. 11. 2017).

57 Vgl. http://www.bbaw.de/telota/software/ediarum (letzter Zugriff: 6. 11. 2017).

58 Zum Vorhaben vgl. http://avhr.bbaw.de/ (letzter Zugriff: 6. 11. 2017).

59 Vgl. https://de.wikisource.org (letzter Zugriff: 6. 11. 2017).

60 Vgl. auch die DTAE-Projektseite: http://www.deutschestextarchiv.de/dtae/ (letzter Zugriff: 6. 11. 2017).

aus größeren Textsammlungen und verschiedenen Formaten (semi-)automatisch in das DTABf zu konvertieren. Andererseits werden Wissenschaftler und Wissenschaftlerinnen angehalten und geschult, ihre individuell erstellten Daten gemäß den etablierten DTA-Vorgaben aufzubereiten und idealerweise über die DTA-Plattform auch weiteren Forscherinnen und Forschern zur Verfügung zu stellen. Über DTAE wurden bereits in über zwanzig Projekten Daten aus sehr verschiedenen Quellen und unterschiedlichen Formaten kuratiert, nicht allein größere Quellenkorpora wie die Editionstexte des Vorhabens *Johann Friedrich Blumenbach – online*[61] oder ein Korpus aus 151 Texten der Wikisource,[62] sondern auch kleinere, individuelle Datensammlungen (oft in älteren Formaten), wie etwa *Texte der deutschen Frauenbewegung*, digitalisiert von Anna Pfundt,[63] ein Korpus von Fach- und Gebrauchstexten, digitalisiert an der JLU Gießen (Thomas Gloning),[64] sowie ein historisches Zeitungskorpus, digitalisiert von Michel Lefèvre (Lefèvre 2013; in Bearbeitung).

Ein Angebot in diesem Zusammenhang ist z. B. die Möglichkeit des verteilten Arbeitens und des Crowdsourcing in DTAQ. Hier ist es möglich, als Arbeitsgruppe die eigenen Projektdaten in einem versionierten System zu verbessern und seitenbasiert tiefer zu annotieren oder auch andere Interessierte zur Mithilfe einzuladen. Für die Vorbereitung von Texten für die Integration in DTAQ wird ein Framework für den Autor-Modus des oXygen-XML-Editors bereitgestellt, das die Bearbeitung von DTABf-konformen Annotationen in einer WYSIWYG-Ansicht ermöglicht. Ein Webformular erleichtert die Erfassung von Metadaten nach DTABf. Regelmäßige Schulungen und Workshops geben Anleitung für die Anwendung der genannten Tools und die Arbeit mit dem DTA.

Die nach DTABf aufbereiteten Daten finden aus dem DTA mit automatisierten Verfahren ihren direkten Weg in die CLARIN-Infrastruktur: über CMDI-Metadatensätze im VLO, über die Anbindung aller DTA-Daten an die Föderierte Volltextsuche, über den TCF-Download sowie über die Vergabe persistenter Identifizierer und die Aufnahme in das CLARIN-Repositorium der BBAW.[65] Auf

---

61 Vgl. die Projektseite: http://www.blumenbach-online.de/ (letzter Zugriff: 6. 11. 2017) sowie die Beschreibung unter http://www.deutschestextarchiv.de/doku/textquellen#blumenbach (letzter Zugriff: 6. 11. 2017).

62 Vgl. die Projektseite: https://de.wikisource.org (letzter Zugriff: 6. 11. 2017) sowie die Beschreibung unter: http://www.deutschestextarchiv.de/doku/textquellen#wikisource (letzter Zugriff: 6. 11. 2017).

63 Vgl. http://www.deutschestextarchiv.de/search/metadata?corpus=tdef (letzter Zugriff: 6. 11. 2017).

64 Vgl. http://www.deutschestextarchiv.de/doku/clarin_kupro_liste?g=tg (letzter Zugriff: 6. 11. 2017).

65 https://clarin.bbaw.de/ bzw. https://clarin.bbaw.de/de/repo/ (letzter Zugriff: 6. 11. 2017).

diese Weise wird die Nachnutzung, die Nachhaltigkeit und längerfristige Archivierung der digitalen Quellen ohne größeren Aufwand für die Datengeber gewährleistet.

Die Integration von Texten und Textsammlungen aus unterschiedlichen Quellen in das DTA trägt somit zu deren besserer Verfügbarkeit und Dissemination bei. Zudem werden die betreffenden Werke über die Plattform miteinander vernetzt, sodass sie im Zusammenhang und in Bezug zueinander recherchierbar und auswertbar sind. Dadurch können Beziehungen zwischen Werken verifiziert werden oder auch erstmals zutage treten.

Zudem konnten aus Texten unterschiedlicher Quellen Spezialkorpora gebildet werden, die das DTA-Kernkorpus sinnvoll ergänzen. So wurden im DTA Zeitungen aus bislang fünf verschiedenen Sammlungen zusammengeführt: die *Neue Rheinische Zeitung*,[66] das *Mannheimer Korpus historischer Zeitungen*,[67] der *Hamburgische Correspondent*,[68] Zeitungen aus dem Korpus der *diGiTexte*[69] sowie die Zeitschriften *Die Grenzboten* (1841–1922),[70] und J. G. Dinglers *Polytechnisches Journal* (1820–1931).[71] Des Weiteren konnte ein Korpus von Texten Alexander von Humboldts aus verschiedenen Quellen zusammengebracht werden: Humboldts

66 Digitalisiert am Rande des Vorhabens *Marx-Engels-Gesamtausgabe* (http://mega.bbaw.de/projektbeschreibung [letzter Zugriff: 6. 11. 2017]); zugänglich unter: http://www.deutschestextarchiv.de/doku/nrhz (letzter Zugriff: 6. 11. 2017).

67 Digitalisiert am Institut für Deutsche Sprache Mannheim (http://www1.ids-mannheim.de/ [letzter Zugriff: 6. 11. 2017]); zugänglich unter: http://www.deutschestextarchiv.de/search/metadata?corpus=mkhz (letzter Zugriff: 6. 11. 2017).

68 Digitalisiert im Rahmen des Projekts *Volltextdigitalisierung der Staats- und Gelehrte[n] Zeitung des Hamburgischen Unpartheyischen Correspondenten und ihrer Vorläufer (1712–1848)* unter der Leitung von Prof. Dr. Britt-Marie Schuster an der Universität Paderborn in Zusammenarbeit mit dem DTA (https://kw.uni-paderborn.de/institut-fuer-germanistik-und-vergleichende-literaturwissenschaft/germanistische-und-allgemeine-sprachwissenschaft/schuster/forschung/projekte/der-hamburgische-unpartheyische-correspondent-volltextdigitalisierung/ [letzter Zugriff: 6. 11. 2017]). Zugänglich unter: http://www.deutschestextarchiv.de/search/metadata?corpus=correspondent (letzter Zugriff: 6. 11. 2017).

69 Digitalisiert in verschiedenen Projektkontexten an der Justus-Liebig-Universität (JLU) Gießen, Professur für Germanistische Sprachwissenschaft (Schwerpunkt Sprachverwendung), Prof. Dr. Thomas Gloning. Zugänglich unter: http://www.deutschestextarchiv.de/doku/clarin_kupro_liste?g=tg (letzter Zugriff: 6. 11. 2017).

70 Digitalisiert im Rahmen eines DFG-Projekts der Staats- und Universitätsbibliothek (SuUB) Bremen zur Nachbearbeitung des OCR-Volltextes der Zeitschrift *Die Grenzboten*, http://brema.suub.uni-bremen.de/grenzboten (letzter Zugriff: 6. 11. 2017). Recherchierbar unter: http://kaskade.dwds.de/dstar/grenzboten/ (letzter Zugriff: 6. 11. 2017).

71 Digitalisiert im Rahmen des DFG-Projekts *Dingler Online* an der Humboldt-Universität zu Berlin (zugänglich unter: http://www.polytechnischesjournal.de/ [letzter Zugriff: 6. 11. 2017]). Recherchierbar unter: http://kaskade.dwds.de/dstar/dingler/ (letzter Zugriff: 6. 11. 2017).

unselbständige Schriften,[72] Nachschriften zu seinen Kosmos-Vorlesungen,[73] ausgewählte gedruckte Werke[74] und perspektivisch Briefe und Reisetagebücher Humboldts aus dem Projekt *Alexander von Humboldt auf Reisen*.[75] Darüber hinaus konnte eine umfangreiche Sammlung von Funeralschriften in das DTA integriert werden. Diese setzt sich aus Funeralschriften der ehemaligen Stadtbibliothek Breslau,[76] aus Epicedien Simon Dachs[77] sowie aus individuell digitalisierten Texten[78] zusammen.

Dass die geschilderten Bemühungen um die Korpora des DTA auch für die wissenschaftliche Community relevant sind, zeigt sich an der zunehmenden wissenschaftlichen Wahrnehmung und Nutzung der Korpora. So werden Ressourcen des DTA im Rahmen von Einzelstudien herangezogen oder empfohlen (z. B. Gloning 2016; Seim 2016; Schuster 2017). Darüber hinaus bauen mehrere eigene Forschungsprojekte wesentlich auf den Daten oder der Infrastruktur des DTA auf.[79]

---

72 Digitalisiert im Rahmen des Projekts *Digitalisierung ausgewählter unselbstständiger Schriften Alexander von Humboldts*, das vom DTA in Kooperation mit dem Vorhaben der BBAW *Alexander von Humboldt auf Reisen* (http://www.bbaw.de/forschung/avh-r [letzter Zugriff: 6. 11. 2017]) und der Professur für Romanische Literaturwissenschaft der Universität Potsdam (Prof. Dr. Otmar Ette) durchgeführt wurde. Zugänglich unter: http://www.deutschestextarchiv.de/search/metadata?corpus=avh (letzter Zugriff: 6. 11. 2017).

73 Digitalisiert im Rahmen einer Kooperation des DTA mit dem Projekt *Hidden Kosmos* (siehe auch oben, Abschnitt 2.4). Zugänglich unter: http://www.deutschestextarchiv.de/search/metadata?corpus=avhkv (letzter Zugriff: 6. 11. 2017).

74 Digitalisiert für das DTA-Kernkorpus; zugänglich unter: http://www.deutschestextarchiv.de/api/pnd/118554700 (letzter Zugriff: 6. 11. 2017).

75 In Vorbereitung, vgl. http://edition-humboldt.de/ (letzter Zugriff: 24. 11. 2017).

76 Digitalisiert im Rahmen des DFG-Projekts *AEDit Frühe Neuzeit* der Herzog August Bibliothek Wolfenbüttel (HAB), des DTA an der BBAW und der Forschungsstelle für Personalschriften an der Philipps-Universität Marburg (http://diglib.hab.de/?link=029 [letzter Zugriff: 6. 11. 2017]). Zugänglich unter: http://www.deutschestextarchiv.de/search/metadata?corpus=aedit (letzter Zugriff: 6. 11. 2017).

77 Digitalisiert im Rahmen des DFG-Pilotprojektes zum *OCR-Einsatz bei der Digitalisierung der Funeralschriften der Staatsbibliothek zu Berlin*; Federbusch & Polzin 2013; zugänglich unter: http://www.deutschestextarchiv.de/search/metadata?corpus=sbb_funeralschriften (letzter Zugriff: 6. 11. 2017).

78 Digitalisiert im Rahmen von Projektarbeiten im Rahmen von Seminaren zu Methoden der digitalen Textedition an der Freien Universität zu Berlin (Dozenten: Matthias Boenig, Susanne Haaf).

79 Aktuell sind hierbei z. B. die Projekte *Syntaktische Grundstrukturen des Neuhochdeutschen. Zur grammatischen Fundierung eines Referenzkorpus Neuhochdeutsch* (http://gepris.dfg.de/gepris/projekt/279165027 [letzter Zugriff: 6. 11. 2017]); *Redewiedergabe – Eine literatur- und sprachwissenschafliche Korpusanalyse* (http://gepris.dfg.de/gepris/projekt/322751860) und *Digitale Sammlung Deutscher Kolonialismus* (bewilligt im März 2017) zu nennen.

5 Zusammenfassung und Ausblick

Das Deutsche Textarchiv wurde von 2007 bis 2016 von der Deutschen Forschungsgemeinschaft gefördert. In dieser Zeit wurde ein nach Textsorten und über die Zeit ausgewogenes Kernkorpus von Texten aus der Zeit von etwa 1600 bis 1900 im Umfang von ca. 120 Millionen Textwörtern aufgebaut. Darüber hinaus entwickelte sich das DTA zu einer von vielen Textproduzenten genutzten Korpusplattform. Zwischen 2011 und 2016 konnten Kooperationen mit über zwanzig institutionell geförderten Korpusprojekten vereinbart werden. Hierdurch entstanden zahlreiche weitere Texte für das DTA. Insgesamt stehen somit im DTA insgesamt mehr als 1 Million Seiten hochqualitativer deutschsprachiger Texte zur Verfügung. Das DTA hat darüber hinaus eine Vielzahl von Nachnutzungen erfahren. Allein zwischen Juli 2015 und April 2017 wurde das DTA als „Gesamtpaket" mehr als 5.000-mal heruntergeladen und zu Forschungszwecken oder für die universitäre Lehre eingesetzt. Etwa 50 Nachnutzungen des Korpus führten zu Veröffentlichungen im Kontext der Digital Humanities.[80]

Im Dezember 2016 lief die Förderung des Deutschen Textarchivs durch die DFG aus. Durch die Aufnahme in den CLARIN-Verbund seit 2014 können technische Softwareentwicklungen und auch der Aufbau eigener Korpora zwar nicht mehr in dem Maße wahrgenommenen werden wie in der durch die DFG geförderten Aufbauphase. Seinen zentralen Aufgaben kann das DTA jedoch durch die Integration in das CLARIN-Zentrum weiter nachkommen. Die BBAW ist seit 2016 im deutschen Teil von CLARIN, dem Projekt CLARIN-D, Koordinator des Kompetenzbereichs *Historische Daten* und kann in diesem Rahmen zentralen Aufgaben der Nutzung des Deutschen Textarchivs nachkommen. Diese bestehen im weiteren zuverlässigen Betrieb der Plattformen DTA und DTAQ, der Weiterentwicklung bzw. Wartung des interoperablen TEI-kompatiblen Schemas DTABf sowie in der nachhaltigen Aufbewahrung der Korpusressourcen und der Services des DTA. Im Rahmen von CLARIN-D engagiert sich das DTA nach wie vor, um durch Kooperationen mit Korpusaufbauprojekten die Textbasis des DTA zu vergrößern. Zusammengefasst kann somit festgestellt werden, dass das DTA das selbstgesteckte Ziel, als aktives Archiv zu fungieren, erfüllt und damit als wichtiger Bestandteil im „Ökosystem" der Digital Humanities verankert ist.

80 Vgl. http://www.deutschestextarchiv.de/clarin-kooperationen (letzter Zugriff: 6. 11. 2017).

Literatur

Bird, Christian, Tim Menzies & Thomas Zimmermann (2015): *The art and science of analyzing software data*. San Francisco, CA: Morgan Kaufmann.

DFG (2015a): *Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora*. Hrsg. vom Fachkollegium Sprachwissenschaften der Deutschen Forschungsgemeinschaft (DFG). http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf (letzter Zugriff: 6. 11. 2017).

DFG (2015b): *Förderkriterien für wissenschaftliche Editionen in der Literaturwissenschaft*. Hrsg. vom Fachkollegium Literaturwissenschaft der DFG. http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/foerderkriterien_editionen_literaturwissenschaft.pdf (letzter Zugriff: 6. 11. 2017).

Dumont, Stefan & Martin Fechner (2014/15): Bridging the gap: Greater usability for TEI encoding. *Journal of the Text Encoding Initiative [Online]* 8. http://jtei.revues.org/1242 (letzter Zugriff: 6. 11. 2017).

Dumont, Stefan, Susanne Haaf, Tobias Kraft, Alexander Czmiel, Christian Thomas & Matthias Boenig (2016): Applying standard formats and tools, 69 f. *TEI Conference and Members Meeting 2016: Book of abstracts*. http://tei2016.acdh.oeaw.ac.at/sites/default/files/TEIconf2016_BookOfAbstracts.pdf#page=71 (letzter Zugriff: 6. 11. 2017).

Eckart, Kerstin (2012): Resource annotations. Aspects of annotations. *CLARIN-D User Guide*. Chapter I,3[,1]. http://media.dwds.de/clarin/userguide/text/annotation_aspects.xhtml.

Federbusch, Maria & Christian Polzin (2013): *Volltext via OCR – Möglichkeiten und Grenzen. Testszenarien zu den Funeralschriften der Staatsbibliothek zu Berlin – Preußischer Kulturbesitz. Mit einem Erfahrungsbericht von Thomas Stäcker aus dem Projekt „Helmstedter Drucke Online“ der Herzog August Bibliothek Wolfenbüttel*. Berlin: o. V. (= Beiträge aus der Staatsbibliothek zu Berlin – Preußischer Kulturbesitz 43.).

Geyken, Alexander, Susanne Haaf, Bryan Jurish, Matthias Schulz, Christian Thomas & Frank Wiegand (2012): TEI und Textkorpora: Fehlerklassifikation und Qualitätskontrolle vor, während und nach der Texterfassung im Deutschen Textarchiv. *Jahrbuch für Computerphilologie – online*. http://computerphilologie.digital-humanities.de/jg09/geykenetal.html (letzter Zugriff: 6. 11. 2017).

Geyken Alexander, Susanne Haaf & Frank Wiegand (2012): The DTA ‘base format’. A TEI-subset for the compilation of interoperable corpora. In Jeremy Jancsary (Hrsg.), *Proceedings of the 11th Conference on Natural Language Processing (KONVENS)*, 383–391. Vienna: Eigenverlag ÖGAI (= Schriftenreihe der Österreichischen Gesellschaft für Artificial Intelligence 5). http://www.oegai.at/konvens2012/proceedings.pdf#page=383 (letzter Zugriff: 6. 11. 2017).

Geyken, Alexander, Matthias Boenig, Susanne Haaf, Bryan Jurish, Christian Thomas, Frank Wiegand & Kay-Michael Würzner (2015): Zeitliche Verlaufskurven in den DTA- und DWDS-Korpora: Wörter und Wortverbindungen über 400 Jahre (1600–2000). In *DHd 2015: Von Daten zu Erkenntnissen: Book of Abstracts*, 78–88. http://gams.uni-graz.at/o:dhd2015.abstracts-vortraege#page=78 (letzter Zugriff: 6. 11. 2017).

Gloning, Thomas (2016): Kommunikationsgeschichte, Themengeschichte, Ideengeschichte. Beispiele für Zusammenhänge und Lehrszenarien. In Volker Harm, Holger Runow & Leevke Schiwek (Hrsg.), *Sprachgeschichte des Deutschen. Positionierungen in Forschung, Studium, Unterricht*, 181–201. Stuttgart: Hirzel.

Haaf, Susanne (2017): Das DTA-Basisformat in neuem Gewand. In *Im Zentrum Sprache. Untersuchungen zur deutschen Sprache*, 3. März 2017. https://sprache.hypotheses.org/147 (letzter Zugriff: 6. 11. 2017).

Haaf, Susanne & Matthias Schulz (2014): Historical newspapers & journals for the DTA. In: *Language Resources and Technologies for Processing and Linking Historical Documents and Archives – Deploying Linked Open Data in Cultural Heritage (LRT4HDA). Proceedings of the workshop, held at the 9th LREC, May 26–31, Reykjavik (Iceland)*, 50–54. http://www.lrec-conf.org/proceedings/lrec2014/workshops/LREC2014Workshop-LRT4HDA%20Proceedings.pdf#page=57 (letzter Zugriff: 6. 11. 2017).

Haaf, Susanne, Alexander Geyken & Frank Wiegand (2014/2015): The DTA 'base format': A TEI subset for the compilation of a large reference corpus of printed text from multiple sources. *Journal of the Text Encoding Initiative 8*. http://jtei.revues.org/1114 (letzter Zugriff: 6. 11. 2017).

Haaf, Susanne & Christian Thomas (2016): Die Historischen Korpora des Deutschen Textarchivs als Grundlage für sprachgeschichtliche Forschungen. In Volker Harm, Holger Runow & Leevke Schiwek (Hrsg.), *Sprachgeschichte des Deutschen. Positionierungen in Forschung, Studium, Unterricht*, 217–234. Stuttgart: Hirzel.

Haaf, Susanne & Christian Thomas (2016 [2017]): Enabling the encoding of manuscripts within the DTABf: Extension and modularization of the format. *Journal of the Text Encoding Initiative (jTEI) 10: Conference Issue*. DOI: 10.4000/jtei.1650. http://jtei.revues.org/1650 (letzter Zugriff: 6. 11. 2017).

Hahn, Udo, Franz Matthies, Erik Faessler & Johannes Hellrich (2016): Uima-based JCoRe 2.0 goes GitHub and Maven Central: State-of-the-art software resource engineering and distribution of NLP pipelines. *Proceedings of the 10th LREC 2016*, 2502–2509. http://www.lrecconf.org/proceedings/lrec2016/pdf/774_Paper.pdf (letzter Zugriff: 6. 11. 2017).

Hamp, Birgit & Helmut Feldweg (1997): GermaNet – a lexical-semantic net for German. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, 9–15. Madrid: o. V. http://www.aclweb.org/anthology/W97-0802 (letzter Zugriff: 6. 11. 2017).

Heid, Ulrich, Helmut Schmid, Kerstin Eckart & Erhard Hinrichs (2010): A corpus representation format for linguistic web services: The D-SPIN text corpus format and its relationship with ISO standards. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC '10)*. Malta: o. V.

Hellrich, Johannes, Franz Matthies & Udo Hahn (2017): UIMA als Plattform für die nachhaltige Software-Entwicklung in den Digital Humanities. In *Dhd 2017: Digitale Nachhaltigkeit: Book of Abstracts*, 279–281. http://www.dhd2017.ch/wp-content/uploads/2017/03/Abstractband_def3_M%C3%A4rz.pdf (letzter Zugriff: 24. 11. 2007).

Henrich, Verena & Erhard Hinrichs (2010). GernEdiT – The GermaNet editing tool. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC '10)*, 2228–2235. Malta: o. V. http://www.lrec-conf.org/proceedings/lrec2010/pdf/264_Paper.pdf (letzter Zugriff: 6. 11. 2017).

Hinrichs, Erhard, Marie Hinrichs & Thomas Zastrow (2010): WebLicht. Web-based LRT services for German. In *Proceedings of the ACL 2010 System Demonstrations*, 25–29.

Jurish, Bryan (2012): Finite-state canonicalization techniques for historical German. Dissertation. Potsdam: Universität Potsdam. urn:nbn:de:kobv:517-opus-55789. http://opus.kobv.de/ubp/volltexte/2012/5578/ (letzter Zugriff: 6. 11. 2017).

Jurish, Bryan & Kay-Michael Würzner (2013): Word and sentence tokenization with Hidden Markov Models. *Journal for Language Technology and Computational Linguistics* 28(2), 61–83.

Jurish, Bryan, Christian Thomas & Frank Wiegand (2014): Querying the Deutsches Textarchiv. In U. Kruschwitz, F. Hopfgartner, & C. Gurrin (Hrsg.), *Proceedings of the Workshop MindTheGap 2014: Beyond Single-Shot Text Queries: Bridging the Gap(s) between Research Communities*, 25–30. urn:nbn:de:0074-1131-6. http://ceur-ws.org/Vol-1131/ (letzter Zugriff: 24. 11. 2017).

Jurish, Bryan, Alexander Geyken & Thomas Werneke (2016): DiaCollo: diachronen Kollokationen auf der Spur. In *DHd 2016: Modellierung – Vernetzung – Visualisierung: Book of Abstracts*, 172–175. Duisburg: nisaba verlag.

Jurish, Bryan (2015): DiaCollo: On the trail of diachronic collocations. In K. De Smedt (Hrsg.), *Proceedings of the CLARIN Annual Conference 2015*, 28–31. http://www.deutschestextarchiv.de/files/jurish2015diacollo-clarin.pdf (letzter Zugriff: 6. 11. 2017).

Kuczera, Andreas (2017): Das Deutsche Textarchiv in der Graphenwelt. In *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte*. https://mittelalter.hypotheses.org/10025 (letzter Zugriff: 6. 11. 2017).

Lefèvre, Michel (2013): *Textgestaltung, Äußerungsstruktur und Syntax in deutschen Zeitungen des 17. Jahrhunderts. Zwischen barocker Polyphonie und solistischem Journalismus*. Berlin: Weidler.

Lemnitzer, Lothar, Bryan Jurish und Daniel Burkhardt (2016): *DiaCollo Tutorial*. http://kaskade.dwds.de/diacollo-tutorial (letzter Zugriff: 6. 11. 2017).

Seim, Stefanie (2016): *Nominalphrasen in literarischen Texten: Strukturtypen und Funktionen beim Figurenentwurf in Werken des 20. und 21. Jahrhunderts*. Gießen: Gießener Elektronische Bibliothek (= Linguistische Untersuchungen 10).

Sinclair, Stéfan & Geoffrey Rockwell (2017): *Voyant Tools*. http://voyant-tools.org/ (letzter Zugriff: 6. 11. 2017).

Thomas, Christian & Frank Wiegand (2015): Making great work even better. Appraisal and digital curation of widely dispersed electronic textual resources (c. 15th–19th centuries) in CLARIN-D. In Jost Gippert & Ralf Gehrke (Hrsg.), *Historical corpora. Challenges and perspectives*, 181–196. Tübingen: Narr.

Tiepmar, Jochen, Thomas Eckart, Dirk Goldhahn & Christoph Kuras (2016): Canonical text services in CLARIN – Reaching out to the Digital Classics and beyond. In *CLARIN Annual Conference*. http://cts.informatik.uni-leipzig.de/documents/CLARIN_CTS.pdf (letzter Zugriff: 6. 11. 2017).

Tiepmar, Jochen, Thomas Eckart, Dirk Goldhahn & Christoph Kuras (2017): Integrating canonical text services into CLARIN's search infrastructure. In: *Linguistics and Literature Studies* 5, 99–104. doi:10.13189/lls.2017.050205, http://www.hrpub.org/journals/article_info.php?aid=5738 (letzter Zugriff: 6. 11. 2017).

Wittenburg, Peter & Dieter van Uytvanck (2012): Metadata. The Component Metadata Initiative (CMDI). In *CLARIN-D User Guide*, Chapter I,2[,6] sowie dies. Metadata. Aggregation. Ebd., Chapter I,2[,7]. http://media.dwds.de/clarin/userguide/text/metadata_aggregation.xhtml (letzter Zugriff: 6. 11. 2017).

Andrea Rapp

11 Digitale Forschungsinfrastrukturen für die Germanistische Mediävistik

allen händen, die sich zum anbau dieses feldes anschicken, ist vollauf arbeit zugedacht

Abstract: Digitale Forschungsinfrastrukturen haben auch in der Germanistischen Mediävistik einen Paradigmenwechsel bewirkt: Sowohl durch den ubiquitären und transparenten Zugriff auf unikale Objekte als auch durch die neuen Forschungs- und Vernetzungsmöglichkeiten für Materialien wie auch für Forschende. Der Beitrag gibt einen Einblick in die Entwicklung digitaler Forschungsinfrastrukturen für die Mediävistik, die teilweise zu den vorbildhaften Modell- und Pionierleistungen in diesem Feld zu gelten haben. Ausgehend von Überlegungen zum Verhältnis von historischer Sprachwissenschaft und moderner Linguistik werden Begriffe und Konzepte digitaler Forschungsinfrastrukturen konturiert. Nach einem kurzen Blick auf die Bemühungen um digitale Angebote seit Beginn der Computernutzung in den Geisteswissenschaften in den späten Vierzigerjahren des 20. Jahrhunderts werden aktuelle Forschungsinfrastrukturen vorgestellt. Zu diesem Zweck werden die Angebote in die Bereiche „Quellen und Nachweisinstrumente" – also die für die Mediävistik besonders wichtige Handschriftendigitalisierung und -erschließung –, „Korpora und Editionen" mit einem Schwerpunkt auf linguistische Korpora, Wörterbücher – vorrangig zu den historischen Sprachstufen – und „Sprachatlanten, Tools und Werkzeuge zur Erschließung und Analyse" sowie „Fachkommunikation" gegliedert. Eine zentrale Aufgabe für die Zukunft ist die Sicherung der Nachhaltigkeit digitaler Angebote; es wird weiterhin darum gehen, die Balance zu finden zwischen Weiterentwicklung und Bewahrung sowie Stabilität und Dynamik.

Keywords: Digitalisierung, Fachkommunikation, Forschungswerkzeuge, Handschriften, Wörterbücher

Anmerkung: Letzter Zugriff für alle im Beitrag genannten URLs: 15. 9. 2017.

Andrea Rapp, Technische Universität Darmstadt, Institut für Sprach- und Literaturwissenschaft, Dolivostraße 15, D-64293 Darmstadt, E-Mail: rapp@linglit.tu-darmstadt.de

https://doi.org/10.1515/9783110538663-012

1 Sprachgeschichte im Kontext des Instituts für deutsche Sprache

Das Institut für deutsche Sprache in Mannheim (IDS) hat als zentrale außeruniversitäre Einrichtung die Aufgabe, die deutsche Gegenwartssprache und ihre neuere Geschichte zu dokumentieren und zu erforschen. Ist damit ein dedizierter Forschungsauftrag für die ältere, mittelalterliche Sprachgeschichte zunächst nicht direkt gegeben, ist die Erforschung und Dokumentation der historischen Entwicklung und des Sprachwandels dennoch ein zentraler Schlüssel zum Verständnis und zur Beschreibung von System und Gebrauch rezenter Sprachzustände. Hermann Paul formulierte das 1880 in seinen *Principien der Sprachgeschichte* mit einem gewissen Absolutheitsanspruch:

> Es ist eingewendet worden, dass es noch eine andere wissenschaftliche Betrachtung der Sprache gäbe als die geschichtliche. Ich muss das in Abrede stellen. Was man für eine nichtgeschichtliche Betrachtung der Sprache erklärt, ist im Grunde nichts als eine unvollkommene geschichtliche, unvollkommen teils durch Schuld des Betrachters, teils durch Schuld des Beobachtungsmaterials. Sobald man über das bloße Konstatieren von Einzelheiten hinausgeht, sobald man versucht, den Zusammenhang zu erfassen, die Erscheinungen zu begreifen, so betritt man auch den geschichtlichen Boden, wenn auch vielleicht ohne sich klar darüber zu sein. (Paul 1880: § 10)

In der modernen Linguistik stehen synchrone und diachrone Sprachbetrachtung nebeneinander und ergänzen einander. Neben dem Eigenwert mediävistischer Forschung und mediävistischen Erkenntnisinteresses bleibt die Rolle im Kontext gegenwartssprachlicher Forschung und Vermittlung. Daher ist auch entgegen ihrer häufigen Vernachlässigung in schulischen und universitären Curricula der „Überblick über die Geschichte der deutschen Sprache", ihre „historische Dimension" Bestandteil des gymnasialen Lehrplans, der die Vermittlung des Sprachsystems als „dem historischen Wandel unterworfen" thematisiert und einfordert (Ländergemeinsame inhaltliche Anforderungen 2008: 23). In den Forschungsschwerpunkten von Ludwig M. Eichinger tritt die Sprachgeschichte ganz selbstverständlich neben Syntax und Wortbildung des Deutschen, Regionalsprachforschung, Soziolinguistik, Stilistik, Textlinguistik und Wissenschaftsgeschichte. Als Direktor einer außeruniversitären Einrichtung, die der Leibniz-Gemeinschaft angehört, übernimmt er zudem besondere Verantwortung für die digitale Forschungsinfrastruktur. Im Jahr 2016 wurde u. a. dieser Bereich im IDS zu einer eigenen Einheit als Programmbereich etabliert, der „die sich mit der nationalen und internationalen Vernetzung der IT-bezogenen Aktivitäten des Hauses und der Sicherung der Nachhaltigkeit von sprachbezogenen Forschungsdaten" bündelt (Eichinger 2016: 5).

Im Folgenden soll ein Einblick in die Vielfalt der Angebote und der Perspektiven gegeben werden, die digitale Technologien für die Erforschung und Vermittlung von Kultur, Sprache und Literatur des Mittelalters bereithalten. Dabei liegt ein Fokus naturgemäß auf Angeboten aus Deutschland bzw. dem deutschsprachigen Raum, obwohl Angebote für die Germanistische Mediävistik nicht auf diesen Raum beschränkt, sondern vielmehr global und weltweit vernetzt sind. Nach einer übergreifenden definitorischen Näherung sollen digitale Forschungsinfrastrukturen für die Germanistische Mediävistik vorgestellt werden, dabei kann es aufgrund der Fülle des Angebots, das zudem ständig weiter ausgebaut wird, nicht um einen vollständigen Überblick gehen, vielmehr sollen einige exemplarisch herausgegriffen und ihre Rolle in Forschung und Lehre adressiert werden. Übergreifende Komponenten – wiewohl entscheidende Grundlage nachhaltiger digitaler Infrastrukturen wie *Persistent Identifier*, Ontologien, *Linked Data* und generische Umgebungen und Initiativen wie TextGrid, CLARIN, DARIAH u. a. seien im Rahmen dieses Beitrags um der Konzentration auf die Mediävistik willen ausgeschlossen. Da es insbesondere für die Epoche des Mittelalters oder für die Vorgeschichte des Deutschen, für die Zeit und Kultur der Germanen zahllose Laienangebote, häufig sogar unseriöse und fragwürdige Angebote im Netz zu finden sind, ist es umso wichtiger, dass die zünftige Mediävistik wissenschaftliche Angebote digital bereitstellt sowie Kriterien der Qualitätssicherung formuliert und kommuniziert.[1] Auch dafür ist eine Institution wie das IDS immens wichtig. Im Anschluss an den Überblick sollen Perspektiven und Desiderata unter den besonderen Herausforderungen von Nachhaltigkeit und Weiterentwicklung angesprochen werden.

2 Definition Forschungsinfrastrukturen

Das Wort Infrastruktur (‚Unter-bau' aus lateinisch infra ‚unterhalb' und structura ‚Zusammenfügung') hat eine vergleichsweise junge Geschichte und umfasst nach der Definition im *Wörterbuch der deutschen Gegenwartssprache* (WDG) „alle für die Wirtschaft eines Landes notwendigen Einrichtungen und

1 Nicht jedes Angebot von Laien (Nicht-Wissenschaftlerinnen und Nicht-Wissenschaftler) ist unwissenschaftlich oder von schlechter Qualität. Gerade für eine stark Klischee-behaftete Epoche wie das Mittelalter, die auf breites Interesse stößt, würden sich *Citizen Science*-Ansätze anbieten. Zur Mittelalterfaszination und Laienrezeption siehe die Aktivitäten des Arbeitskreises *Mittelalterrezeption* um die Germanisten Stefan Keppler-Tasaki (Tokyo/Berlin) und Mathias Herweg (Karlsruhe), z. B. Keppler-Tasaki & Herweg (2012), Rohr (2011), zur Präsenz des Mittelalters in Sprache, Kultur und Gesellschaft vgl. z. B. auch die Einleitung in Fuhrmann (1996, 1987: 15–16).

Anlagen, die nur mittelbar der Produktion dienen: zur I. einer Volkswirtschaft gehören Eisenbahnen, Straßen, Häfen".[2] Auch die Belege im IDS-Wortinformationssystem OWID sowie das Kookkurrenzprofil aus den IDS-Corpora belegen dieses Verständnis von Infrastruktur als Hauptbedeutung.[3] Der Begriff „Forschungsinfrastruktur" wurde dementsprechend zunächst vor allem auf wissenschaftliche Großgeräte in den Naturwissenschaften wie Teilchenbeschleuniger, Teleskope oder Labore bezogen, während erst in den letzten 10 bis 15 Jahren zum einen auch die geisteswissenschaftlichen Äquivalente wie Bibliotheken, Archive oder Museen unter diesem Blickwinkel betrachtet wurden (Wissenschaftsrat Bericht 2017: 7–9 & Empfehlungen 2011) und zum anderen ein Verständnis digitaler Forschungsinfrastrukturen entwickelt wurde und dadurch insgesamt eine Ablösung von der einseitigen Vorstellung von Infrastruktur als Großgerät erfolgte. Die Relevanz digitaler Infrastruktur zeigt sich auch in der Einrichtung des Rats für Informationsinfrastrukturen, der seit 2014 Politik und Wissenschaft in strategischen Zukunftsfragen der digitalen Wissenschaft berät.[4]

Forschung in den Geisteswissenschaften basierte seit jeher auf Infrastrukturen. Im analogen Bereich gehören dazu Artefakte, Bücher, Bibliotheken, Archive oder Museen. Im digitalen Bereich arbeiten wir mit ihren digitalen Derivaten als „Daten", mit Korpora und Verzeichnissen, Wörterbüchern und Informationssystemen, vor allem aber auch mit digitalen Werkzeugen wie Recherche-, Editions-, Auswertungs- oder Statistiktools. Die umfassende digitale Transformation erforderte und erfordert große Anstrengungen seitens der Wissenschaft und der Politik und zwingt zu einer Revision, aber auch zu einer notwendigen Reflexion geisteswissenschaftlichen Arbeitens und Forschens. Diese Reflexion muss Gewinne und Verluste betrachten, schließlich zu einer neuen Art geisteswissenschaftlichen Forschens führen, in der analoge Arbeitsweisen nicht disruptiv von digitalen abgelöst werden, sondern in der durch das Zusammengehen beider Ansätze unsere Kenntnisse erweitert werden. Der digitale Wandel und der Aufbau neuer Infrastrukturen ist unbestritten zu einem wichtigen Motor geisteswissenschaftlichen Forschens geworden. Er führt auch zu einer veränderten Wahrnehmung der Geisteswissenschaften in anderen Wissenschaftsbereichen und in der Gesellschaft, was gerade auch in der aktuellen gesellschaftlichen und politischen Diskussion kaum hoch genug

---

2 Laut Verlaufskurve im *Digitalen Wörterbuch der deutschen Gegenwartssprache* ist das Wort seit den 1920er Jahren präsent: http://zwei.dwds.de/wb/Infrastruktur

3 www.owid.de/artikel/202070; http://corpora.ids-mannheim.de/ccdb/?preload=http://corpora.ids-mannheim.de/ccdb/db/496e/496e667261737472756b747572/t496e667261737472756b74757200.html?src=elex

4 Koalitionsvertrag zwischen CDU, CSU und SPD, 18. Legislaturperiode, Dezember 2013.

geschätzt werden kann. Die Relevanz und die Chancen digitaler Infrastrukturen wurden in verschiedenen Papieren des Wissenschaftsrats betont:

> Forschungsinfrastrukturen leisten in allen Wissenschaftsbereichen wesentliche Beiträge zum wissenschaftlichen Erkenntnisgewinn, zur wissenschaftlichen Beantwortung von Fragen gesellschaftlicher Relevanz sowie zur internationalen Anschlussfähigkeit dieser Anstrengungen. [...] sie wandeln sich von tradierenden und Fachinformationen bevorratenden Hilfseinrichtungen zu Inkubatoren für neue und innovative wissenschaftliche Fragestellungen aufgrund von Forschungsdaten, die durch diese Infrastrukturen selbst erst erzeugt werden. [...] Digital aufbereitete Fachinformationen bieten durch ihre Verknüpfung mit Metadaten ganz neuartige Möglichkeiten der forschenden Erschließung von Bibliotheks-, Archiv- und Sammlungsbeständen. (Wissenschaftsrat Empfehlungen 2011: 7)

Trotz aller Erfolge und der Breite des Angebots bleiben die Geisteswissenschaften in Aufbau und Nutzung, d. h. auch bei den Forschungsmöglichkeiten, hinter anderen Wissenschaftszweigen und in Förderprogrammen zurück, sie haben immer noch Nachholbedarf. Eine große Herausforderung liegt derzeit zudem in der Konzeption und Umsetzung des nachhaltigen Betriebs digitaler Infrastrukturen (Neuroth & Rapp 2016) – eine für die Geisteswissenschaften und das kulturelle Erbe besonders wichtige Anforderung –, was nicht allein eine technologische Aufgabe ist, sondern vor allem eine organisatorische und politische Herausforderung darstellt. Außeruniversitären Institutionen kommt in diesem Zusammenhang eine wichtige Rolle zu, insbesondere, wenn sie auch mit universitärer Forschung gut vernetzt sind und einen klaren inhaltlichen und gesellschaftlich akzeptierten Auftrag haben. Die soziale Dimension von Forschungsinfrastruktur (Wissenschaftsrat 2011: 10) ist ferner für das eher kleine Fachgebiet der Germanistischen Mediävistik, das auf Vernetzung in hohem Maße angewiesen ist, besonders relevant.

Für die (Germanistische) Mediävistik ist das Potenzial digitaler Zugänglichkeit evident, nicht zuletzt wegen der Gebundenheit der Texte an unikale Überlieferung, und man war in weiten Teilen der digitalen Transformation zumindest aufgeschlossen, z. T. wurde auch die Entwicklung führend (mit-)gestaltet. Eine der ersten geisteswissenschaftlich motivierten Nutzungen des Computers war Roberto Busas in den Vierzigerjahren des 20. Jahrhunderts startende und heute noch digital zugängliche und genutzte corpus- und computerlinguistische Aufarbeitung der Werke und der Sprache des Thomas von Aquin[5], ein Projekt, an dem beispielhaft die gesamte Entwicklung der *Digital Humanities* „erzählt" werden kann (Nyhan & Terras 2017; Terras, Nyhan & Vanhoutte 2013); Busa gilt daher als Gründungsvater der *Digital Humanities*. Indices und

---

5 www.corpusthomisticum.org

Konkordanzen sowie die Erstellung von Editionen waren ab den 1960er und 1970er Jahren Schwerpunkte;[6] das von Wilhelm Ott und anderen entwickelte philologische Werkzeug TUSTEP ermöglichte die Umsetzung zahlloser Vorhaben und Forschungen und ist ebenfalls bis heute im Einsatz.[7] Beides sicher ermutigende Beispiele digitaler Langlebigkeit, getragen durch die Interessen der Nutzenden. Die 1990er Jahre standen unter der Vision der „Verteilten Nationalen Forschungsbibliothek", die von der DFG u. a. mit der Einrichtung zweier Digitalisierungszentren in München und Göttingen sowie der Neuordnung der Förderprogramme im heutigen Programmbereich LIS verbunden waren (Göttker 2016). Im Bereich der Handschriftendigitalisierung ist u. a. auf die Universitätsbibliothek Graz zu verweisen, wo neben der Digitalisierung des Katalogs auch der Grazer Büchertisch für die Digitalisierung mittelalterlicher Handschriften entwickelt wurde, der sich – zusammen mit den entsprechenden Verfahren und Workflows – rasch zum Standard entwickelte.[8] Die ebenfalls in den 1990er Jahren erfolgte Retro-Digitalisierung und Vernetzung der mittelhochdeutschen Wörterbücher gehörte zu den Pionierprojekten der Volltextdigitalisierung und setzte Maßstäbe für die TEI/XML-Kodierung und damit strukturell-semantische Erschließung und Vernetzung von Wörterbuchdaten (Burch et al. 2003). Dennoch ist das Potenzial all dieser (Einzel-)Leistungen und Ergebnisse noch kaum ansatzweise ausgeschöpft. Vor allem in der Vernetzung und nachhaltigen Erschließung ist noch viel zu tun, damit zukünftig Linked-Data-Annotationen ein offenes und stabiles Netz weben können, das dem von Hans Walter Gabler für Editionen skizzierten „web of discourses" (Gabler 2010: 46) entspricht. Wie Luise Borek (2017) in ihrer Dissertation resümiert, wäre das Ergebnis eines solchen koordinierten Vernetzungsprozesses „dann kein annotiertes Corpus, sondern ein dynamisches und interaktives Wissensnetz, das auch Diskurse einschließt", und sie benennt weiter die zentrale Voraussetzung für ein solches Netz:

---

6 In Europa und für die Germanistische Mediävistik hat Roy Wisbey hier innovativ gewirkt; siehe z. B. EADH People, Terras 2014; in Deutschland ist die Reihe *Maschinelle Verarbeitung altdeutscher Texte* (I–V), basierend auf einer Reihe internationaler Tagungen zwischen 1971 und 1997, bei denen auch Vertreter des IDS präsent waren, anzuführen.

7 Tübinger System von Textverarbeitungs-Programmen (TUSTEP), www.tustep.uni-tuebingen.de. Die offen zugänglichen Protokolle des Kolloquiums über die Anwendung der EDV in den Geisteswissenschaften an der Universität Tübingen sind eine hervorragende Quelle zur Entwicklung des Computereinsatzes in den Geisteswissenschaften: www.tustep.uni-tuebingen.de/kolloq.html

8 „1993/94 entstand die erste Homepage der Sondersammlungen, wenig später war der Handschriften-Katalog der Universitätsbibliothek Graz der erste Online-Katalog in Österreich im Altbuch-Bereich", konnte Hans Zotter zu Recht stolz berichten. „1995 wurden die ersten Schritte in der Digitalisierung gesetzt." (Presseaussendung Uni Graz 2009).

> Der Zugriff auf die darin enthaltenen Daten – wie auch auf die allgemein stetig anwachsenden Forschungsdaten – kann nur dann transparent erfolgen, wenn einzelne Ressourcen identifizierbar und in Netzstrukturen eingebunden sind. Die Voraussetzung ist eine flächendeckend tragfähige, d. h. nachhaltige und standardisierte Digital Curation, die nur von entsprechender Infrastruktur zu leisten ist. (Borek 2017: 199)

Probleme bereiten neben Mängeln bei persistenter Identifizierung auch nach wie vor rechtliche Restriktionen, denn viele Ressourcen sind nicht unter offener Lizenz verfügbar oder erst mit mehrjährigen *Moving Walls*, was sich nachteilig auf die Forschung auswirkt. Dieses Problem betrifft also nicht allein die Sprachdaten der Neuzeit, sondern auch die Objekte, Ressourcen und Forschungsergebnisse zur Sprachgeschichte.

3 Digitale Forschungsinfrastrukturen in der (Germanistischen) Mediävistik – ein kursorischer Überblick

Um die unterschiedlichen Herausforderungen und Funktionalitäten und den heterogenen Erschließungsstand adressieren zu können, ist der folgende Überblick gegliedert in die Bereiche Quellen und Nachweisinstrumente, Korpora und Editionen, Wörterbücher und Atlanten, Werkzeuge sowie Fachkommunikation.

Die Technologien und Workflows zur Erfassung und Erschließung (im Sinne von Rohdaten) sind mittlerweile gut etabliert, standardisiert und dokumentiert (Jannidis, Kohle & Rehbein 2017) und haben vor allem mit den DFG-Praxisregeln und weiteren Handreichungen breite Wirkung entfaltet (DFG Praxisregeln Digitalisierung 2013; DFG Handreichungen 2013; DFG Förderkriterien 2015). Die *Koordinierte Förderinitiative zur Weiterentwicklung von Verfahren für die Optical-Character-Recognition (OCR)* von 2014 mit dem entsprechenden Koordinierungsprojekt OCR-D[9] wird die Verfügbarmachung historischer Sprachdaten entscheidend verbessern, nachdem bislang vor allem Double-Keying-Verfahren gute Qualität für heterogene, nicht-normierte Sprachdaten leisteten.

Auf der Seite der Formate und Standards müssen die Schlagwörter Unicode und XML/TEI fallen:[10] 1987 wurde die *Text Encoding Initiative* (TEI) mit Unter-

9 http://ocr-d.de/

10 Es sei in diesem Zusammenhang darauf hingewiesen, dass neben dem für die Langfristarchivierung unverzichtbaren XML andere Technologien leistungsfähige Möglichkeiten zur Recherche und Analyse von Texten bieten, wie z. B. die Graphtechnologien; dazu z. B. Kuczera (2015).

stützung der *Association for Computers and the Humanities*, der *Association for Computational Linguistics* und der *Association for Literary and Linguistic Computing* gegründet. Dieses Konsortium ist wissenschaftsgetrieben und diskutiert und entwickelt hard- und softwareunabhängige Methoden für die Annotation, die Interoperabilität und die langfristige Archivierung von Textdaten; sie bieten als Metasprache den Vorteil, dass sie wie natürliche Sprachen mit einem begrenzten Inventar und grammatischen Regeln quasi unendlich viele Dinge formulieren und formalisieren können. Damit sind neben textstrukturellen Annotationen und Metadaten auch linguistische Erschließungen und semantische Annotation in beliebiger Tiefe und Komplexität möglich.

3.1 Quellen und Nachweisinstrumente

Digitalisate unikaler (bzw. begrenzt zugänglicher) Überlieferung bieten viele Vorteile: Sie ermöglichen den direkten Zugang der Forschung zu den Quellen „bei gleichzeitiger Schonung der kostbaren, bisweilen fragilen Originale“ (DFG Praxisregeln Digitalisierung 2013: 5). Hierhin gehören auch ingenieurwissenschaftliche Leistungen wie der „Grazer Büchertisch“ (Mayer 1999), der eine berührungsfreie und besonders schonende Aufnahme erlaubt und der in zahlreichen Institutionen weltweit im Einsatz ist.

Mit den 1963 erstmals erschienenen DFG-Richtlinien zur Katalogisierung (^{5}1992) wurde Deutschland weltweit führend in der Erschließung mittelalterlicher Handschriften. Die Katalogisierungskompetenz wird heute in sechs Handschriftenzentren[11] gebündelt. Als technische Plattform und Portal steht derzeit noch *Manuscripta Mediaevalia* zur Verfügung, das Katalogisate von aktuell „90.000 Dokumente[n] zu abendländischen Handschriften hauptsächlich in deutschen Bibliotheken“[12] offen recherchierbar macht. Es erlaubt die gezielte Recherche nach mittelalterlichen Handschriften in deutschen Bibliotheken und Archiven und enthält darüber hinaus zahlreiche digitalisierte Handschriftenkataloge und Links zu Handschriftendigitalisaten. Neuere Katalogisierungs- und Handschriftendigitalisierungsprojekte lieferten standardmäßig ihre Daten an das Portal, so dass das Angebot laufend erweitert und vervollständigt wurde.

Ein Programm zur Digitalisierung aller verfügbaren Handschriften gibt es in Deutschland noch nicht, doch wird aktuell im Rahmen von DFG-geförderten Pilotprojekten daran gearbeitet, die methodischen, technischen und organisa-

---

11 www.handschriftenzentren.de
12 Selbstdarstellung auf: www.manuscripta-mediaevalia.de

torischen Grundlagen „für großflächige Digitalisierungsmaßnahmen“[13] zu formulieren und auch *Manuscripta Mediaevalia* neu zu konzipieren – in dieser Übergangsphase werden von der DFG keine Förderanträge zur Handschriftendigitalisierung entgegengenommen.[14] Dennoch gibt es national wie international eine Vielzahl von Initiativen und Projekten, mit denen Bibliotheken und Archive ihre Bestände digital ins Netz bringen. Hier haben nicht zuletzt die Katastrophen von Weimar und Köln das Bewusstsein für die Möglichkeiten und Notwendigkeiten digitaler Archivierung und Bestandserhaltung geschärft. Vorreiter waren Projekte wie das zur vollständigen Digitalisierung der Dom- und Diözesanbibliothek in Köln.[15] Die Arbeiten und Initiativen der Bibliotheken von St. Gallen, Heidelberg, München, Trier, Wolfenbüttel und vielen anderen erweitern das Spektrum der Möglichkeiten und treiben die technischen und wissenschaftlichen Entwicklungen auf diesem Feld kontinuierlich voran. Da es in diesem Bereich also eine besonders dynamische Entwicklung gibt, ist der oben angesprochene Masterplan von größter Bedeutung. Um ansatzweise den Überblick über aktuelle Digitalisierungen zu behalten, sei auch auf die entsprechenden Aufstellungen auf Portalen wie mediaevum[16] und archivportal-d[17] sowie auf das Zentralverzeichnis digitalisierter Drucke ZVDD[18] verwiesen (dazu unten mehr).

Für deutschsprachige Handschriften des Mittelalters existiert mit dem *Marburger Handschriftencensus*[19] eine weitere zentrale Anlaufstelle, deren Pflege und Ausbau seit 2016 durch ein langfristiges Akademievorhaben gesichert ist. Das vom Institut für deutsche Philologie des Mittelalters der Philipps-Universität Marburg und nunmehr von der Akademie der Wissenschaften und der Literatur (Mainz) betreute Portal geht im Kern auf zwei Repertorien zurück, auf das *Repertorium deutschsprachiger Handschriften des 13. und 14. Jahrhunderts* und das *Paderborner Repertorium der deutschsprachigen Textüberlieferung des 8. bis 12. Jahrhunderts*, die die entsprechenden Bestände erschließen. Im Akademievorhaben werden Neufunde aufgenommen, unvollständige Angaben ergänzt

---

13 www.handschriftenzentren.de/wp-content/uploads/2016/06/Priorisierungsfragen-Masterplan_pub.pdf

14 www.handschriftenzentren.de/wp-content/uploads/2016/06/Priorisierungsfragen-Masterplan_pub.pdf; s. auch die Informationen zum Stand des Antrags zu einem neuen Handschriftenportal, das Manuscripta Mediaevalia ablösen soll, unter www.handschriftenzentren.de/handschriftenportal

15 www.ceec.uni-koeln.de/

16 www.mediaevum.de

17 www.archivportal-d.de

18 www.zvdd.de

19 http://handschriftencensus.de/

sowie vor allem die Handschriften ab dem 14. Jahrhundert erschlossen, die den Großteil mittelalterlicher Überlieferung ausmachen. Ziel ist demnach die vollständige Verzeichnung deutschsprachiger mittelalterlicher Handschriften. Suchmöglichkeiten, übersichtliche und standardisierte Darbietungen der Beschreibungen sowie aktuelle Links zu Digitalisaten machen das Repertorium zu einem hervorragenden Arbeitsmittel, das auch für Studierende geeignet ist. Die Akademieförderung erlaubt nun auch wieder die Bereitstellung des Meldedienstes für neue Handschriftenfunde, Ergänzungen und Korrekturen, die stark nachgefragt, jedoch aus Ressourcengründen eingestellt worden war – wiederum zeigt sich zum einen die Relevanz dieser Infrastrukturen für die wissenschaftliche Kommunikation und Vernetzung, zum anderen die dringende Notwendigkeit der Bindung an langfristige Institutionen.

3.2 Korpora und Editionen

Aktuelle Editionsprojekte werden in der Regel digital erstellt und beinhalten neben der Druckedition zumeist auch eine digitale Publikationskomponente, während ältere, aber immer noch gültige und wertvolle Editionen nachträglich retrodigitalisiert werden mussten und müssen. Aufgrund der Fülle des Angebots können hier nur exemplarisch einige recht bekannte umfangreiche Sammlungen, Initiativen und Mustereditionen vorgestellt werden, die frei im Netz verfügbar sind.

Das von der Universität Trier und der University of Virginia, Charlottesville, USA, bereits 1999 durchgeführte Pionier-Projekt des *Mittelhochdeutschen Textarchivs* stellt rund 100 mittelhochdeutsche Texte in höchster Erfassungsqualität zur Verfügung, viele davon als freie XML-Downloads.[20] Andere sind auch über das Quellenarchiv des *Mittelhochdeutschen Wörterbuchs* zugänglich (siehe unten), denn das Vorhaben sollte vor allem das Belegkorpus für die Erstellung des neuen Mittelhochdeutschen Wörterbuchs digital verfügbar machen. Grundlage des Textarchivs waren daher Standardeditionen der Texte, die im Double-Keying-Verfahren erfasst wurden und in Kooperation mit dem damals in diesem Bereich führenden eText-Center der University Library in Charlottesville unternommen wurde, was den Know-how-Transfer in beide Richtungen enorm beförderte.

Unter dem Dach *Deutsch Diachron Digital* (DDD) wurden balancierte Referenzkorpora zu den älteren Sprachstufen des Deutschen erstellt, die in einem sprachstufenübergreifenden tiefenannotierten Korpus historischer Texte des

20 http://mhgta.uni-trier.de/

Deutschen gebündelt sind. Das *Referenzkorpus Altdeutsch*[21] erfasst und annotiert die ältesten Sprachdenkmäler des hochdeutschen und des niederdeutschen Sprachraumes vom Beginn der schriftlichen Überlieferung um 750 bis etwa 1050 (Humboldt-Universität zu Berlin, Goethe-Universität Frankfurt am Main, Friedrich-Schiller-Universität Jena). Annotiert werden mindestens Satzgrenzen, Wortarten und Morphologie, tiefere Annotationen sind ebenfalls möglich. Analog verfährt das *Referenzkorpus Mittelhochdeutsch* (1050–1350) der Universitäten Bonn und Bochum, so dass die Korpora zu einem späteren Zeitpunkt zusammengefügt werden können. Aufgrund der Tiefenannotation können differenzierte und komplexe linguistische Abfragen durchgeführt werden, so dass das DDD-Korpus das Potenzial hat, zu einem zentralen Forschungsinstrument für die historische Linguistik zu werden und Anschluss an Korpora neuerer Sprachstufen verspricht. Anders als bei Referenzkorpora zu rezenten Sprachstufen wie das am IDS betreute,[22] das ständig erweitert werden und Milliarden von laufenden Wortformen umfassen kann, sind die Korpora älterer Sprachstufen naturgemäß begrenzt und von Überlieferungszufällen abhängig – ein Umstand, der deutlicher zu Tage tritt, je weiter man in die Überlieferungsgeschichte zurückgeht. Die DDD-Korpora zum Mittelhochdeutschen und Frühneuhochdeutschen sind nach verschiedenen Varietätenaspekten soweit als möglich balanciert zusammengestellt und mit Metadaten und Annotationen versehen, um die Recherche- und Auswertungsergebnisse kritisch einordnen zu können (Lemnitzer & Zinsmeister 2010: 44–46, Perkuhn; Keibel & Kupietz 2012: 47–48).

Ein lange etabliertes Angebot zur allgemeinen historischen und vergleichenden Sprachwissenschaft liegt vor mit dem *Thesaurus Indogermanischer Text- und Sprachmaterialien* (TITUS),[23] die Jost Gippert (Frankfurt a. M.) aufgebaut hat. Darunter ist Etliches zu älteren Kleinsprachen bzw. Sprachen mit schmaler Überlieferung, so dass hier Wichtiges z. B. zum Cimbrischen versammelt ist; dabei wird den konkreten Überlieferungszeugnissen ebenso Beachtung geschenkt wie der Transkription, der kritischen Edition und der grammatischen und semantischen Erschließung mit Hilfe digitaler Werkzeuge.

Die Editionsphilologie gehört zu den *early adopters* digitaler Technologien und setzte den Computer früh als Werkzeug, nach der Etablierung des *World Wide Web* auch als Medium für kritische Editionen ein (Gärtner 2011).[24] Als

21 www.deutschdiachrondigital.de/

22 www1.ids-mannheim.de/kl/projekte/korpora/

23 http://titus.fkidg1.uni-frankfurt.de/framed.htm?/index.htm

24 Vgl. dazu auch die Tagungen und Publikationen der AG Germanistische Editionen: www.ag-edition.org/

moderne Edition, die sich die Möglichkeiten des digitalen Mediums gut nutzbar macht, sei die *Parzival-Edition*[25] von Michael Stolz und seinem Team von der Universität Bern herausgegriffen. Neben Handschriftendigitalisaten mit entsprechenden Transkriptionen enthält diese Präsentation Editionsproben ausgewählter Abschnitte, die sich durch Parallelpräsentationen und transparente Apparatgestaltung auszeichnen. In diesem Feld sind zwei Trends zu beobachten: Zum einen werden Kanontexte – z. T. solche mit komplexer Überlieferung wie der *Parzival* – digital neu ediert, zum anderen auch Texte, die eher zu den *big unread* zu zählen sind, digital verfügbar gemacht und ausgewertet.

3.3 Wörterbücher und Sprachatlanten

Einen Überblick über die wissenschaftlichen Wörterbücher, die zumeist von den deutschen Wissenschaftsakademien erstellt wurden und werden, gibt das Wörterbuchportal, das von der Berlin-Brandenburgischen sowie der Heidelberger Akademie der Wissenschaften betreut wird.[26] Es beinhaltet nicht nur eine Linkliste zu den entsprechenden Vorhaben, sondern auch eine übergreifende Suchmöglichkeit über verschiedene Volltextwörterbücher.

Das von Kurt Gärtner, Klaus Grubmüller und Karl Stackmann begründete neue *Mittelhochdeutsche Wörterbuch* wird vollständig digital erarbeitet und mit einer moderaten *moving wall* von sechs Monaten nach Erscheinen der jeweiligen Lieferung frei im Netz publiziert.[27] Das von den Akademien in Göttingen und Mainz betreute Vorhaben mit Arbeitsstellen in Göttingen, Mainz und Trier hat seit 2014 Ludwig M. Eichinger als Projektleiter. Seine digitale Publikation ist nicht allein mit dem digitalen Quellenarchiv, sondern auch mit den älteren Wörterbüchern verknüpft, so dass die Hilfsmittel zum Mittelhochdeutschen hervorragend zugänglich sind. Erfreulich ist, dass auch die laufenden Vorhaben zum Althochdeutschen[28] und Frühneuhochdeutschen[29] mittlerweile digital zugänglich sind. Zwar ist eine differenzierte übergreifende Recherche aufgrund der heterogenen Anlage der Wörterbücher nicht einfach, dennoch ist hier ein weiterer Schritt zu einem alle Sprachstadien umfassenden Wörterbuch des Deutschen getan – zu weitergehenden Erschließungs- und Vernetzungsmöglichkeiten unten mehr.

25 www.parzival.unibe.ch/editionen.html
26 www.woerterbuch-portal.de/
27 www.mhdwb-online.de/
28 http://awb.saw-leipzig.de
29 https://fwb-online.de/

Die Wörterbücher zum Althochdeutschen und Frühneuhochdeutschen wurden anders als das von Anfang an digital konzipierte Mittelhochdeutsche Akademie-Wörterbuch im laufenden Betrieb retrodigitalisiert. Hier waren die älteren Wörterbücher zum Mittelhochdeutschen Vorreiter:[30] Die drei aufeinander bezogenen abgeschlossenen Wörterbücher zum Mittelhochdeutschen (Benecke, Müller, Zarncke, Lexer und das *Findebuch zum mittelhochdeutschen Wortschatz*) sind volltextdigitalisiert, untereinander verknüpft und voll recherchierbar sowohl als Verlagsangebot als auch online frei im Netz zugänglich und stehen nicht nur für die Nachschlagearbeit beim Übersetzen, sondern auch als differenziert durchsuchbare Datenbank für linguistische Arbeiten zur Verfügung. Ihre digitale Zugänglichkeit überbrückt zum einen die Zeit bis zur Fertigstellung des *Neuen Akademiewörterbuchs*, zum anderen bieten sie aufgrund ihres breiteren Zeitausschnitts (z. T. bis 1500, in Einzelfällen bis ins 16. Jh.) andere Informationen, die in den modernen Sprachstadienwörterbüchern im mittelhochdeutschen und frühneuhochdeutschen Wörterbuch getrennt sind.

Als ein Spezialwörterbuch sei hier noch das *Deutsche Rechtswörterbuch* erwähnt, das den Wortschatz der Sprache des Rechtslebens vom Beginn der schriftlichen Überlieferung im 5. Jahrhundert in merowingischen Urkunden bis hin zu Goethe erfasst und dabei die westgermanischen Varietäten berücksichtigt.[31] Dank seiner Vernetzung mit anderen digitalen Wörterbüchern, insbesondere aber mit seinem Quellenarchiv, wird das Online-Angebot des DRW zu einem umfassenden Informationssystem ausgebaut.

Aufgrund seines umfassenden Konzepts kann auch das *Deutsche Wörterbuch* der Brüder Grimm für die älteren Sprachstufen herangezogen werden, denn der ursprüngliche Plan sah eine Dokumentation der deutschen Sprache ab ca. 1450 vor.[32] Sehr häufig reichen die Belege und Beschreibungen jedoch darüber hinaus bis ins Alt- und Mittelhochdeutsche zurück bzw. umfassen im Etymologieteil auch die Vorgeschichte.

Auch Dialekt- bzw. regionalsprachliche Wörterbücher liefern wertvolle sprachhistorische Informationen, sowohl was Lexik und Semantik, aber auch Lautwandel (z. B. zum Stand der zweiten Lautverschiebung) betrifft – sie sind damit wichtige Bindeglieder zwischen historischen und rezenten Varietäten und darüber hinaus geeignet, interessante didaktische und öffentlichkeitswirksame Zugänge zu schaffen.

30 www.mwv.uni-trier.de/
31 www.rzuser.uni-heidelberg.de/~cd2/drw/
32 http://dwb.uni-trier.de/de/

Das Trierer Wörterbuchnetz führt die genannten und weitere verschiedene Wörterbücher unter einer gemeinsamen Suchoberfläche zusammen und vernetzt Artikelteile semantisch.[33] In diesem großartigen Werkzeug steckt noch viel Potenzial für eine umfassende, auf transparenten Kriterien beruhende systematische lexikographische Vernetzung und Forschungen zu linguistischen wie auch infrastrukturell-technologischen Aspekten.

Wie erwähnt, kann auch über die Dialektologie ein Zugang zu sprachhistorischen Themen geschaffen werden, so dass hier auch auf den *Digitalen Wenkeratlas* (DiWA)[34] hingewiesen werden muss. Wie der *Wenkeratlas* selbst stellt auch die digitale Version, die erstmals das Potenzial dieses Dokumentations- und Forschungswerkzeuges auszuschöpfen vermag, eine veritable Pionierleistung in der digitalen Sprachkartographie und den Digital Humanities dar. Neben Wenkers gedruckten und auch den bislang nur in zwei handgezeichneten Exemplaren vorhandenen Karten sind weitere Regionalatlanten, die Digitalisate der Original-Wenkerbögen, Tondokumente sowie eine Bibliographie angebunden. Leicht zu bedienende Werkzeuge ermöglichen Überblendungen der Karten, was Einblicke in Dynamik und Variation regionalsprachlicher Erscheinungen bietet. Damit ist dieses Werkzeug nicht allein für die Forschung, sondern auch für die Lehre und die Vermittlung im curricularen Seminar, im Schulunterricht oder auch an Laien von größtem Wert.

3.4 Werkzeuge

Noch ein Wort zu digitalen Spezialwerkzeugen jenseits linguistischer Werkzeuge, Korpustools und *Natural Language Processing*: Handschriftendigitalisate erlauben viel mehr als den „lesenden" Zugriff des menschlichen Auges, denn an die Digitalisate lassen sich bestimmte Forschungsfragen richten, die mit entsprechenden digitalen Werkzeugen bearbeitet werden können. Sehr gut etabliert ist die digitale Paläographie in verschiedenen Domänen und Disziplinen, die sich in erster Linie um automatische Zeichen- und Handschriftenerkennung sowie Digitale Paläographien und Verzeichnisse bemüht (Ciula 2005; *DigiPal*; Quirke 2011; Stokes 2014; Fecker, Märgner & Schaßan 2015). Mittelalterliche Handschriften, insbesondere die *codices picturati* des *Sachsenspiegels*, waren Gegenstand automatischer Analyse und Deutung von Bildmustern (Yarlagadda et al. 2013; SemToNotes). Die automatische Layoutanalyse gehört zu den Grundlagen im Bereich der OCR-Systeme, die vor allem bei moderneren Druckwerken, sehr großen Zeitschriftkorpora u. a. m. erfolgreich eingesetzt

33 www.woerterbuchnetz.de
34 www.diwa.info

wird. (Mittelalterliche) Handschriften weisen gegenüber den gedruckten Vorlagen eine individuelle Variation auf, die größere Herausforderungen an die Automatisierung stellt. Im Projekt *eCodicology* konnte der Nachweis geführt werden, dass Untersuchungen zu physikalischen Eigenschaften und zur materiellen Gestalt von Handschriften am Digitalisat durchgeführt werden können (Busch et al. 2017).[35] In den Möglichkeiten, die für die Auswertung digitaler Image-Derivate mittelalterlicher Zeugnisse aufgezeigt werden, liegt ein großes Zukunftspotenzial, das noch kaum angedacht, geschweige denn ausgeschöpft ist. Die Übertragbarkeit auf andere Materialen, z. B. Briefsammlungen und Zeitschriftenkorpora ist ebenfalls bereits erwiesen.

3.5 Fachkommunikation

Seit vielen Jahren hat sich das gut betreute und gepflegte, als Privatinitiative und mittlerweile in eine GbR überführte mediävistische Fachportal *mediaevum.de*[36] etabliert, das nicht nur als zuverlässiger Wegweiser zu fachspezifischen Angeboten wie digitalen Publikationen, Hilfsmitteln und Datenbanken dient, sondern auch Aufgaben der Kommunikation innerhalb der Fachcommunity übernimmt: Dazu gehören Funktionalitäten wie die von der Community selbst gepflegte Projektdatenbank, ein Tagungskalender, ein Stellenmarkt oder ein Diskussionsforum mit zahlreichen, auch für Anfänger interessanten Themenknoten.

Auf dem Blogportal *Hypotheses*[37] bündelt eine Initiative von Martin Bauch, Karoline Döring und Björn Gebert ein vielfältiges und professionelles Angebot zur interdisziplinären Wissens- und Wissenschaftskommunikation. Klassische Informationsformate wie Tagungsüberblick, Rezensionsaggregation oder Diskussionsforum werden ergänzt durch verschiedene Publikationsformate, so dass das Portal eine echte Lücke im wissenschaftlichen Kommunikationsangebot zu schließen vermag.

Eine Brücke zwischen Mittelalter und dem Social-Media-Kanal Twitter (und auch auf Facebook) schlagen seit 2016 Wernfried Hofmeister und Ylva Schwinghammer mit ihrer erfolgreichen Ausstellung #Dichterleben – mittelalterliche *tweets* aus der Steiermark.[38] Sie zeigen damit, wie die Vermittlung mediävis-

---

35 www.ecodicology.org

36 www.mediaevum.de

37 http://mittelalter.hypotheses.org/

38 www.kommunikation.steiermark.at/cms/beitrag/12472070/29771102.

tischer Inhalte auf innovative Weise und auf eine breitere Öffentlichkeit ausgerichtet bereichert werden kann.[39]

4 Ausblick

Digitale Forschungsinfrastrukturen haben auch in der Germanistischen Mediävistik einen Paradigmenwechsel bewirkt: Sowohl durch den ubiquitären und transparenten Zugriff auf unikale Objekte als auch durch die neuen Forschungs- und Vernetzungsmöglichkeiten für Materialien wie auch für Forschende. Es wird nun gelten, die vorhandenen Mosaiksteine übergreifend zusammenzuführen und an die Angebote zur neueren und zur Gegenwartssprache anzubinden. Die deutsche Sprache ist in verschiedenen Spezialwörterbüchern, Korpora und Sprachatlanten in einer Tiefe und Dichte erschlossen wie wenige andere, digitale Forschungsinfrastrukturen und Technologien erlauben die Vision eines umfassenden Informationsnetzwerks zum Deutschen für eine wieder als Einheit verstandene Digitale Philologie. Die Umsetzung dieser Vision erfordert nicht allein technologische und algorithmische Entwicklungen, sondern auch und vor allem philologische Forschungen und Anstrengungen (vom politischen Willen und Ressourcen ganz zu schweigen) – allen Händen ist also vollauf Arbeit zugedacht. Für die digitalen Infrastrukturen wird es weiterhin darum gehen, die Balance zu finden zwischen Weiterentwicklung und Bewahrung, Stabilität und Dynamik (Neuroth & Rapp 2016).

Literatur

AG Germanistische Editionen: www.ag-edition.org

Althochdeutsches Wörterbuch online: http://awb.saw-leipzig.de

Archivportal-D: www.archivportal-d.de

Ausstellung #Dichterleben im Steiermärkischen Landesarchiv eröffnet: Steirische Dichter des Mittelalters twittern in die Welt hinaus: www.kommunikation.steiermark.at/cms/beitrag/12472070/29771102

Borek, Luise (2017): *Arthurische Pferde als Bedeutungsträger. Eine Fallstudie zu ihrer digitalen Klassifizierung*. Diss. masch. Technische Universität Darmstadt.

Burch, Thomas, Johannes Fournier, Kurt Gärtner & Andrea Rapp (Hrsg.) (2003): *Standards und Methoden der Volltextdigitalisierung. Akten des Kolloquiums vom 8. bis 9. Oktober*

---

39 Die Tweets sind Teil einer breiteren Vermittlungsaktion, zu der auch acht regionale Literaturpfade gehören; s. auch Hofmeister 2015.

an der Universität Trier. Stuttgart: Steiner (Schriften der Akademie der Wissenschaften und der Literatur Mainz, Geistes- und Sozialwiss. Klasse 9).

Busch, Hannah, Franz Fischer & Patrick Sahle (hrsg. unter Mitarbeit von Bernhard Assmann, Philipp Hegel & , Celia Krause (2017): *Kodikologie und Paläographie im Digitalen Zeitalter* 4. Norderstedt: Books on Demand (Schriften des Instituts für Dokumentologie und Editorik 11). www.i-d-e.de/publikationen/schriften/11-kpdz4/. BoD: urn:nbn:de:101:1–201707081765.

Ciula, Ariana (2005): Digital palaeography: Using the digital representation of medieval script to support palaeographic analysis. *Digital Medievalist* 1. www.digitalmedievalist.org/journal/1.1/ciula/

Codices Electronici Eclesiae Coloniensis: www.ceec.uni-koeln.de

Corpus Thomisticum: www.Corpusthomisticum.org

Deutsch Diachron Digital: www.deutschdiachrondigital.de

Deutsches Rechtswörterbuch online: www.rzuser.uni-heidelberg.de/~cd2/drw/

Deutsches Referenzkorpus DeReKo: www1.ids-mannheim.de/kl/projekte/korpora/

Deutsches Wörterbuch der Brüder Grimm online: www.dwb.uni-trier.de

DFG Förderkriterien (2015): *Förderkriterien für wissenschaftliche Editionen in der Literaturwissenschaft 2015*. www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/foerderkriterien_editionen_literaturwissenschaft.pdf

DFG Handreichungen (2013): *Handreichungen: Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora 2013*. www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf

DFG Richtlinien Handschriftenkatalogisierung (1992): *Richtlinien Handschriftenkatalogisierung Deutsche Forschungsgemeinschaft, Unterausschuß für Handschriftenkatalogisierung*. 5. erw. Aufl., Bonn-Bad Godesberg: DFG.

DFG Memorandum Digitalisierung Handschriftenbestände (2016). www.dfg.de/foerderung/programme/infrastruktur/lis/lis_foerderangebote/erschliessung_digitalisierung/#micro7827218

DFG Praxisregeln „Digitalisierung“, Stand 02/2013: www.dfg.de/formulare/12_151/12_151_de.pdf

DigiPal: Digital Resource and Database of Manuscripts, Palaeography and Diplomatic, London, 2011–14. www.digipal.eu/

Digitaler Wenkeratlas: www.diwa.info

Digitales Wörterbuch der deutschen Sprache DWDS: http://zwei.dwds.de/

EADH – European Association for Digital Humanities, People, Roy Wisbey. http://eadh-static.adho.org/people/roy-wisbey.html

Eichinger, Ludwig (2016): Vorwort des Direktors. In *Jahresbericht des Instituts für Deutsche Sprache* 16, 5–7. www1.ids-mannheim.de/fileadmin/org/pdf/IDS_Jahresbericht_2016.pdf

Fecker, Daniel, Volker Märgner & Torsten Schaßan (2015): Vom Zeichen zur Schrift: Mit Mustererkennung zur automatisierten Schreiberhanderkennung in mittelalterlichen und frühneuzeitlichen Handschriften. In Constanze Baum & Thomas Stäcker (Hrsg.), *Grenzen und Möglichkeiten der Digital Humanities*. (= Sonderband der *Zeitschrift für digitale Geisteswissenschaften* 1). doi:10.17175/sb001_008.

Frühneuhochdeutsches Wörterbuch online: http://fwb-online.de

Fuhrmann, Horst (1987): *Einladung ins Mittelalter*. München: C. H. Beck.

Fuhrmann, Horst (1996): *Überall ist Mittelalter*. München: C. H. Beck.

Gabler, Hans Walter (2010): Theorizing the Digital Scholarly Edition. *Literature Compass* 7.2, 43–56, doi:10.1111/j.1741-4113.2009.00675.x.

Gärtner, Kurt (2011): Der Computer als Werkzeug und Medium in der Editionswissenschaft. Ein Rückblick. *editio. Internationales Jahrbuch für Editionswissenschaft* 25, 32–41. doi:https://doi.org/10.1515/9783110239362.32.

Göttker, Susanne (2016): *Literaturversorgung in Deutschland. Von den Sondersammelgebieten zu den Fachinformationsdiensten – Eine Analyse*. Wiesbaden: Dinges & Frick.

Grimm, Jacob (1846): *Über die wechselseitigen Beziehungen und die Verbindung der drei in der Versammlung vertretenen Wissenschaften*. Rede auf dem Germanistentag in Frankfurt a. M. 1846, zuerst veröffentlicht in den Verhandlungen der Germanisten Frankfurt am Main 1847; auch in J.G. *Kleine Schriften*, Bd. 7, 1884, 556–563.

Handschriftencensus. Eine Bestandsaufnahme der handschriftlichen Überlieferung deutschsprachiger Texte des Mittelalters: http://handschriftencensus.de

Handschriftenzentren – zentrale Kompetenzeinrichtungen für Handschriftenerschließung und -digitalisierung: www.handschriftenzentren.de

Hofmeister, Wernfried (Hrsg.) (2015): *Literarische Verortungen*. Graz: Keiper.

Frühneuhochdeutsches Wörterbuch online: https://fwb-online.de

Jannidis, Fotis, Hubertus Kohle & Malte Rehbein (Hrsg.) (2017): *Digital Humanities. Eine Einführung*. Stuttgart: Metzler.

Keppler-Tasaki, Stefan & Mathias Herweg (Hrsg.) (2012): *Rezeptionskulturen. 500 Jahre literarischer Mittelalterrezeption zwischen Kanon und Populärkultur*. Berlin/Boston: de Gruyter (Trends in Medieval Philology 27).

Koordinierungsprojekt OCR-D: http://ocr-d.de/

Kuczera, Andreas (2015): Graphdatenbanken für Historiker. Netzwerke in den Registern der Regesten Kaiser Friedrichs III. mit neo4j und Gephi. *Mittelalter. Interdisziplinäre Forschung und Rezeptionsgeschichte*, 5. Mai 2015. http://mittelalter.hypotheses.org/5995

Ländergemeinsame inhaltliche Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung. Beschluss der Kultusministerkonferenz vom 16. 10. 2008 i. d. F. vom 8. 12. 2008.

Lemnitzer, Lothar & Heike Zinsmeister (2010): *Korpuslinguistik. Eine Einführung*. Tübingen: Narr.

Manuscripta Mediaevalia: www.manuscripta-mediaevalia.de

Mayer, Manfred (1999): *Digitalisierung mittelalterlicher Handschriften an der Universitätsbibliothek Graz*. conservation-us.org

Mediaevum.de – Informationsportal für die deutsche Literatur des Mittelalters: www.mediaevum.de

Mittelalter – Interdisziplinäre Forschung und Rezeptionsgeschichte: http://mittelalter.hypotheses.org/

Mittelhochdeutsches Textarchiv online: http://mhgta.uni-trier.de

Mittelhochdeutsches Wörterbuch online: www.mhdwb-online.de

Mittelhochdeutsche Wörterbücher im Verbund: www.mwv.uni-trier.de

Neuroth, Heike & Andrea Rapp (2016): Nachhaltigkeit von digitalen Forschungsinfrastrukturen. *Bibliothek – Forschung und Praxis* 40(2), 264–270, doi: 10.1515/bfp-2016-0022.

Nyhan, Julianne & Melissa Terras (2017): *Uncovering ‘hidden’ contributions to the history of Digital Humanities: The Index Thomisticus’ female keypunch operators*. Conference Paper Digital Humanities 2017, Montreal Canada, August 8–11, 2017. https://dh2017.adho.org/abstracts/358/358.pdf

Online-Wortschatz-Informationssystem Deutsch: www.owid.de/
Parzival-Projekt: www.parzival.unibe.ch/editionen.html
Paul, Hermann ([1]1880/[5]1920): *Principien der Sprachgeschichte*. Tübingen: Niemeyer.
Perkuhn, Rainer, Holger Keibel & Marc Kupietz (2012): *Korpuslinguistik*. Paderborn: Fink.
Presseaussendung Uni Graz (2009): *Eröffnet – Zentrum für die Erforschung des Buch- und Schrifterbes*. Uni Graz 19. Juni 2009. https://static.uni-graz.at/fileadmin/presse/Archiv/0609/newswww_archiv_detail_023.html.
Quirke, Stephan (2011): Agendas for digital palaeography in an archaeological context: Egypt 1800 BC. In Franz Fischer, Christiane Fritze und Georg Vogeler (Hrsg.), *Kodikologie und Paläographie im digitalen Zeitalter* 2, 279–294. Norderstedt: BOD (Schriften des Instituts für Dokumentologie und Editorik, 3).
Rohr, Christian (Hrsg.) (2011): *Alles heldenhaft, grausam und schmutzig?: Mittelalterrezeption in der Populärkultur*. Berlin et al.: LIT Verlag (Austria: Forschung und Wissenschaft – Geschichte 7).
SemToNotes: Semantic Topological Notes, http://hkikoeln.github.io/SemToNotes/
Stokes, Peter (2014): Describing handwriting – again. In Tal Hassner, Robert Sablatnig, Dominique Stutzmann, Ségolène Tarte, Schloss Dagstuhl (Hrsg.), *Digital palaeography: New machines and old texts: Dagstuhl Reports*, Bd. 4, 127–128. DROPS: Dagstuhl Research Online Publication Server. doi: 10.4230/DagRep.4.7.112.
Terras, Melissa (2014): *Roy Wisbey and literary and linguistic computing*. http://melissaterras.blogspot.de/2014/05/roy-wisbey-and-literary-and-linguistic.html (letzter Zugriff: 24. 5. 2014).
Terras, Melissa, Julianne Nyhan & Edward Vanhoutte (2013): *Defining Digital Humanities. A reader*. Farnham: Ashgate.
Thesaurus Indogermanischer Text- und Sprachmaterialien (TITUS): http://titus.fkidg1.uni-frankfurt.de/framed.htm?/index.htm
TUSTEP = Tübinger System von Textverarbeitungs-Programmen www.tustep.uni-tuebingen.de
Wissenschaftsrat (2011): *Empfehlungen des Wissenschaftsrates zu Forschungsinfrastrukturen in den Geistes- und Sozialwissenschaften*, Drs. 10465-11 28. 1. 2011.
Wissenschaftsrat (2017): *Bericht zur wissenschaftsgeleiteten Bewertung umfangreicher Forschungsinfrastrukturvorhaben für die nationale Roadmap* 2017. www.wissenschaftsrat.de/download/archiv/6410-17.pdf
Wörterbuchnetz, Trier: www.woerterbuchnetz.de
Wörterbuchportal: www.woerterbuchportal.de
Corpus Thomisticum: www.corpusthomisticum.org
eCodicology: www.eCodicology.org
Yarlagadda, Pradeep, Antonio Monroy, Bernd Carque & Björn Ommer (2013): Towards a Computer-Based Understanding of Medieval Images. In *Scientific Computing and Cultural Heritage – Contributions in Computational Humanities*, 89–97. Berlin, Heidelberg: Springer.
Zentralverzeichnis digitalisierter Drucke: www.zvdd.de

Der Titel dieses Aufsatzes wurde entnommen aus: Jacob Grimm: Über die wechselseitigen Beziehungen und die Verbindung der drei in der Versammlung vertretenen Wissenschaften. Rede auf dem Germanistentag in Frankfurt a. M. 1846. Zuerst veröffentlicht in den Verhandlungen der Germanisten. Frankfurt am Main 1847; auch in Jacob Grimm (1884): *Kleinere Schriften*, Bd. 7, 556–563. Berlin: Ferd. Dümmlers Verlagsbuchhandlung Harrwitz und Gossmann. Der Satz lautet vollständig:

Für alle zweige deutscher sprache, dies wort in einer völlig zulässigen weitesten bedeutung genommen, eröffnet sich, je weiter die forschung vorrückt, immer lohnendere aussicht, und allen händen, die sich zum anbau dieses feldes anschicken, ist vollauf arbeit zugedacht.

Martine Dalmas und Roman Schneider

12 Die grammatischen Online-Angebote des IDS aus Sicht der Germanistik im Ausland

Gegenwart und Zukunft

Abstract: Seit Mitte der 1990er Jahre wird am Institut für deutsche Sprache (IDS) in Mannheim erforscht, wie der hochkomplexe Gegenstandsbereich „Grammatik“ unter Ausnutzung digitaler Sprachressourcen und hypertextueller Navigationsstrukturen gleichermaßen wissenschaftlich fundiert und anschaulich vermittelt werden kann. Die grammatischen Online-Informationssysteme des IDS wenden sich nicht allein an Forscher und die interessierte Öffentlichkeit in Deutschland, sondern in gleichem Maße an Germanisten und Deutsch-Lernende in der ganzen Welt. Der vorliegende Beitrag beschreibt die damit verbundenen Hoffnungen und Ansprüche. Daran anschließend thematisiert er praktische Einsatzmöglichkeiten und skizziert die funktionale und inhaltliche Weiterentwicklung der digitalen Grammatik-Angebote.

Keywords: Auslandsgermanistik, Deutsch als Fremdsprache, Grammatik, Hypertext, Informationssysteme

1 Sprachbeschreibung heute: zwischen „Schnee von gestern“ und Zukunftsmusik

Zu den Sorgen und Herausforderungen unserer modernen Gesellschaft, aber auch zu ihren neuen Potenzialen gehören moderne Kommunikationsformen, die im Kontext der Internationalisierung entstanden sind und durch den Gebrauch digitaler Informationstechnologien neue Chancen in der Forschung, der Bildung und der Lehre eröffnen. Parallel zu dieser Entwicklung führt die

Martine Dalmas, Sorbonne Université – Centre de linguistique en Sorbonne (EA 7332), 108, bd Malesherbes, F-75850 Paris Cedex 17, E-Mail: martine.dalmas@sorbonne-universite.fr
Roman Schneider, Abteilung Grammatik, Institut für Deutsche Sprache (IDS), R5 6–13, D-69161 Mannheim, E-Mail: schneider@ids-mannheim.de

https://doi.org/10.1515/9783110538663-013

Internationalisierung der Geisteswissenschaften – über die europäischen Grenzen hinaus – zu Anpassungen, d. h. zwangsläufig zu Engpässen, aber auch zu neuen Sichtweisen. Auch die Einbindung der Nationalsprachen in ein mehrsprachiges Europa und vor allem in eine globalisierte Welt kann Hoffnungen und Ängste aufkommen lassen. So hat die Diskussion über Spracheinstellungen und insbesondere über Einstellungen zur deutschen Sprache seit einigen Jahren an Schärfe gewonnen und ist in den Mittelpunkt der Debatte um die Stellung des Deutschen gerückt. Die „außenstehenden" Germanistinnen und Germanisten, die von Natur aus „zwischen den Sprachen" stehen, beobachten mit großem Interesse den Umgang der deutschen Muttersprachlerinnen und Muttersprachler mit ihrer Sprache sowie ihre sprachpflegerischen Ansprüche und sprachpolitischen Erwartungen.

In dieser Hinsicht rückt das Institut für deutsche Sprache (IDS) in den Vordergrund: Es ist und bleibt ein wichtiger Anhaltspunkt und Indikator, nach innen und nach außen, also sowohl in seinen neuen Forschungsorientierungen als auch in Bezug auf seine Angebote und die bereitgestellten Materialien für Sprachinteressierte. Dass Online-Verfahren manche Wege verkürzen und erleichtern und Menschen einander näher bringen, steht außer Zweifel. Diese Möglichkeit wird – neben Bildungsinstitutionen – immer mehr von Wissenschaftlerinnen und Wissenschaftlern sowie wissenschaftlichen Einrichtungen genutzt, um ihre Arbeitsergebnisse einer breiteren Leserschaft zugänglich zu machen, den wissenschaftlichen Dialog zu fördern und den Kontakt zur Öffentlichkeit herzustellen. Dass das IDS da keine Ausnahme bildet, erfreut die Außenstehenden und steht im Dienst der deutschen Sprache!

Zu den vielfältigen Online-Angeboten des IDS gehören bereits seit Ende der Neunzigerjahre die Plattformen *Grammis* und *ProGr@mm* (Breindl, Schneider & Strecker 2000, Schneider & Schwinn 2014), zwei umfangreiche Informationssysteme, die sich an alle Germanisten und Germanistinnen sowie Sprachinteressierte im In- und Ausland wendet. Gegenstand dieser Systeme ist nicht nur die ausführliche Beschreibung der grammatischen Strukturen des Deutschen, sondern auch die Bereitstellung linguistisch relevanter Dokumentationen und Informationsquellen zur Vertiefung bestimmter Aspekte sowie zu weiterführenden methodischen oder inhaltlichen Fragestellungen.

Wir werden hier auf einige Aspekte dieser grammatischen Online-Angebote eingehen, die für die Arbeit mit und an der deutschen Sprache im nichtdeutschsprachigen Ausland als besonders relevant und hilfreich betrachtet werden können. Unter Berücksichtigung der aktuellen Lage an internationalen Universitäten wird zunächst die Frage nach den Zielgruppen, ihrer Beschaffenheit und ihren Erwartungen aufgeworfen, dann werden die Module der beiden grammatischen Angebote kurz vorgestellt und ihre Nutzungsmöglichkeiten als

interaktives Nachschlageinstrument unter die Lupe genommen. Anschließend gehen wir noch einmal auf die Stellung des grammatischen Wissens und auf die möglichen Zugangsformen ein und formulieren einige Desiderata. Zum Abschluss geben wir einen kurzen Ausblick auf die künftige Entwicklung der grammatischen Onlinesysteme des IDS.

2 Grammatik nach der sogenannten „kommunikativen Wende“

Wir gehen davon aus, dass die Relevanz eines wissenschaftlich motivierten Informationssystems mit der Charakterisierung ihrer Nutzerinnen und Nutzern eng verbunden ist. Entsteht ein solches Angebot, das grammatisches Wissen zugänglich macht und weitere Informationsquellen integriert, an einem Forschungsinstitut wie dem IDS, das 1997 eine dreibändige Grammatik der deutschen Sprache hervorgebracht hat, dann gewinnt der Schulterschluss zwischen Forschung und Lehre an Bedeutung, erweitert die Perspektive, aber gerät gleichzeitig in ein Dilemma: Sowohl die Wahl der Inhalte als auch ihre Gestaltung zwingen die Autoren mitunter zu Einschränkungen, Kürzungen und evtl. auch Vereinfachungen, aber gleichzeitig auch zu Erweiterungen (zwecks Explizierung), die die Wünsche und Erwartungen, aber vor allem auch die Defizite und Bedürfnisse der Anwender berücksichtigen sollen.

Wenn *Grammis* und *Progr@mm* sich an Germanistinnen und Germanisten wenden, dann sowohl an Studierende als auch an Lehrende, d. h. in letzterem Fall auch indirekt an Studierende. Ob Deutsch die Muttersprache ist oder nicht: Ihre Vorkenntnisse in der deutschen Grammatik sind oft eher spärlich. Paradoxerweise scheint die sogenannten „kommunikative Wende“ in den Lehr- und Lernmethoden bei den Mutter- sowie auch den Fremdsprachen in den letzten Jahrzehnten die Kenntnis und Beherrschung der grammatischen Bezüge in Äußerungen und Texten stark in den Hintergrund gestellt bzw. reduziert zu haben. Anstatt, wie man es durch eine nunmehr aktivere Rolle der Lernenden im Unterricht hätte erwarten dürfen, die Kenntnis des Systems zu fördern, sie zu erweitern, indem auch Strukturen der gesprochenen Sprache thematisiert, beschrieben und erlernt werden, stellt man in vielen Ländern immer wieder fest, dass Kenntnis und Beherrschung des Sprachsystems eher nachgelassen haben.

Vor diesem Hintergrund ist das Wecken des Interesses für eine erklärende Beschreibung der grammatischen Strukturen des Deutschen, zumal bei Studierenden mit noch mangelhaften Sprachkenntnissen, eine Herausforderung, die nicht unbedingt leicht zu meistern ist. Da, wo man denken könnte, dass die

Darstellung eines zusammenhängenden und kohärenten Systems, die in den Jahren vor der Universität gefehlt hat, besonders willkommen sein könnte, stößt man bei den Studierenden immer wieder auf Widerstandsformen, die zunächst erklärt, evtl. besprochen und auf jeden Fall überwunden werden müssen.

Diese Herausforderung trifft insofern auch die Lehrkräfte, als es dann in vielen Fällen darum geht, das Wissen zu selektieren und zu adaptieren, bevor es weitergegeben bzw. umgesetzt wird. Darauf sind aber die meisten Lehrkräfte nicht vorbereitet, denn die dazu nötige pädagogische Ausbildung bzw. die erwünschte Weiterbildung fehlen weitgehend. Hinzu kommt, dass die Sprachwissenschaft in den Curricula mancher Länder eher ein Orchideendasein fristet, was sich bis in den Lehrerberuf auf die Motivation für eine eingehende und „aufklärerische" Sprachbeschreibung auswirkt.

Dass die kommunikative(n) Kompetenz(en) aber nicht ohne die Beherrschung der verschiedenen Standards auskommt/auskommen und diese – zumindest teilweise – auch auf eine bewusste Kenntnis von System und Gebrauch zurückgeht, sollte außer Frage stehen. Dies gilt für alle Germanisten und Germanistinnen oder DaF-/DaZ-Lehrenden und Lernenden. Auch auf einem höheren Niveau sind geeignete Nachschlageinstrumente vonnöten, sie müssen allerdings differenziert und zweckmäßig ausgewählt werden. Aus dieser Perspektive werden im Folgenden die Möglichkeiten vorgestellt, die mit *Grammis* und *ProGr@mm* angeboten werden.

3 Ein gelungener Spagat: Wie *Grammis* und *ProGr@mm* konzipiert sind

Sowohl *Grammis* (*Grammatisches Informationssystem*) als auch *ProGr@mm* (*Propädeutische Grammatik*) sind digitale Angebote, die dem Nutzer Zugang zu Fachtexten und Ressourcen zur deutschen Grammatik bieten (vgl. z. B. Storrer 1997; Strecker 1998; Schneider 2004). Sie unterscheiden sich allerdings in ihrer inhaltlichen Gestaltung, denn ihre Ziele und ihr Zielpublikum sind zum Teil andere. Auch wenn sie beide als „hypermediale Nachschlageinstrumente per Mausklick" zu grammatischen Fragen zu betrachten sind, indem sie Erklärungen und Hintergrundwissen bieten, und auch wenn sich beide – zumindest ursprünglich – auf die am IDS entstandene Grammatik der deutschen Sprache (Zifonun et al. 1997) als theoretische Grundlage stützen und deshalb eine gemeinsame Grundstruktur aufweisen, sind sie zum Teil unterschiedlich konzipiert und orientiert. Dies hängt nicht zuletzt damit zusammen, dass *ProGr@mm* später entstanden und explizit didaktisch ausgerichtet ist.

Wir gehen hier zunächst auf die Gesamtstruktur beider Plattformen ein und befassen uns dann näher mit dem Teilbereich, der unterschiedlich konzipiert ist: dem jeweiligen Grammatik-Teil.

Die ursprünglichen Implementierungen von *Grammis* und *ProGr@mm* enthalten jeweils sechs Module, die zu 50 % dieselben sind, so dass von inhaltlichen und funktionalen Überschneidungen zwischen den beiden Plattformen gesprochen werden kann. Gemeinsam sind das *Grammatische Wörterbuch* (aus drei bzw. vier Einzelwörterbüchern bestehend: zu Affixen, zu Präpositionen, zur Verbvalenz und – bei *Grammis* – zu Konnektoren), die *Grammatische Bibliografie* (mit einer feingranular ausgearbeiteten Suche nach Objekt- oder Schlagwörtern und der Möglichkeit, individuelle Merklisten zu kompilieren) und das *Terminologische Wörterbuch* bzw. die *Grammatischen Fachtermini* (mit Definitionen der benutzten Fachausdrücke und Angaben zu äquivalenten Bezeichnungen in anderen Sprachen).

Interessanterweise enthält bereits das relativ frühe *Grammis* zwei Module, die sich an zwei ganz verschiedene Nutzerkreise wenden: Das eine Modul (*Korpusgrammatik*) ermöglicht einen Blick in die Hexenküche der Grammatiker, nämlich in die korpusgestützte Forschung, und beinhaltet technische und methodologische Informationen zum Umgang mit Korpora sowie auch – exemplarisch aufzufassende – Ergebnisse spezieller Untersuchungen, die unter Zuhilfenahme authentischer Textkorpora durchgeführt wurden (Pilotstudien zu Fugen-Elementen, Genitiv-Markierungen und AcI-Konstruktionen bzw. *wie*-Sätze); mittelfristig angestrebt wird eine umfassende, dezidiert korpusgestützt erarbeitete Grammatik des Deutschen (vgl. Bubenhofer, Konopka & Schneider 2014). Das andere Modul (*Grammatik in Fragen und Antworten*), das sich eher an ein breiteres Publikum wendet, befasst sich mit Zweifelsfällen bzw. mit häufig gestellten Fragen zu Schwierigkeiten, für die man in den gängigen Grammatiken nicht unbedingt oder nicht sofort eine Antwort findet (vgl. Konopka 2006; Konopka & Schneider 2012). Neben der Grundfrage zur Bedeutung von „Grammatik" werden hier morphologische, syntaktische, aber zum Teil auch stilistische oder lexikalische Fragestellungen beantwortet: Die Breite der angesprochenen Themen zeugt von einem ebenfalls breit aufgefassten Grammatikbegriff, der nicht nur präskriptiv, sondern auch gebrauchsorientiert ist und Standard und Variation, auch regionale, in Verbindung bringt. Für die Hilfesuchenden, die das berühmt-berüchtigte „Kreuz mit Grammatiken" nicht erleben möchten, ist dieses Modul ein besserer, direkter Weg – vorausgesetzt allerdings, ihr grammatisches Problem gehört zu den hier behandelten Themen bzw. sie sind in der Lage, die dazu passende Frage zu finden.

ProGr@mm, später entwickelt und häufig als E-Learning-Hypertextsystem bezeichnet (vgl. Schwinn 2005), ist, wie oben schon angedeutet, didaktisch

ausgerichtet und enthält deswegen zwei Module, die mit einer spezifischen Lehrsituation zusammenhängen: einerseits ein „kontrastives“ Modul, auf das wir unten noch eingehen werden, und ein interaktives *Forum* als Kommunikationsplattform für Seminarveranstaltungen, das auf angemeldete Teilnehmerinnen und Teilnehmer beschränkt ist. Ebenfalls im System angemeldete Seminarleiter und Seminarleiterinnen ermöglichen diesen den Zugang zum Forum und können auf spezielle Systemfunktionalitäten zurückgreifen, die den Austausch mit Studierenden fördern.

Das Modul *Kontrastiv* geht auf ein vom deutschen Staat finanziertes, internationales Projekt mit dem Namen *EuroGr@mm* zurück, das unter Leitung des IDS zwischen 2007 und 2012 fünf Universitäten in nicht-deutschsprachigen Ländern zusammenführte und sich zum Ziel gesetzt hatte, eine Adaption der im *Grammatischen Grundwissen* vorhandenen Inhalte vor dem Hintergrund der jeweiligen Muttersprachen, ihrer Eigenheiten sowie ihrer grammatikalischen bzw. grammatikografischen Traditionen vorzunehmen (vgl. z. B. Augustin 2009; Dalmas, Fabricius-Hansen & Schwinn 2016). Insofern ist es keine kontrastive Grammatik im engen Sinne, sondern eher ein kontrastiver Blick auf die grammatischen Strukturen des Deutschen. Somit wendet sich *ProGr@mm* sowohl an Lehrkräfte und Studierende mit Deutsch als Muttersprache, als auch an solche, deren Herkunft oder Standort außerhalb des deutschen Sprachraums liegen und deren Zugang zur deutschen Sprache den Bezug auf eine andere Sprache (bewusst oder unbewusst) mit einbezieht.

4 Die unterschiedlichen Einsatzmöglichkeiten der grammatischen Online-Angebote

Die unterschiedliche Ausrichtung der beiden Angebote, die sich an Zielgruppen mit zum Teil stark divergierenden Bedürfnissen wenden, ermöglicht differenzierte Zugriffs- und Nutzungsformen, auf die wir nachfolgend eingehen wollen.

Es können drei Hauptsituationen genannt werden, die zur Nutzung von *Grammis* bzw. *Progr@mm* führen können:

a) das gezielte Nachschlagen und Nachprüfen bestimmter Formen, Gebrauchsweisen oder Definitionen,
b) die Suche nach präzisen und eingehenden Kommentaren zu grammatischen und pragmatischen Funktionen, sowie
c) das Zusammenstellen einschlägiger Literatur zu bestimmten Themen.

a. Beim Nachschlagen geht es den Nutzerinnen und Nutzern meistens um ein schnelles Nachprüfen einer Form im Zusammenhang mit ihrer Funktion (bzw. ihren Funktionen) oder um eine Definition. Die Suchoptionen sowie die Querverweise durch Links ermöglichen einen relativ einfachen Zugang aus verschiedenen Perspektiven, der viel leichter und schneller erfolgen kann als ein mühsames Blättern in mehreren Bänden (mit entsprechendem Vormerken von Kapitelüberschriften und Seitenzahlen!). Das unter Heranziehung exemplarischer Phänomene gestaltete Modul *Grammatik in Fragen und Antworten* kommt insbesondere bei grammatischen Zweifelsfällen ins Spiel. Darüber hinaus empfiehlt sich bei Bedarf nach einer gleichermaßen präzisen und prägnanten Begriffsbestimmung der Rückgriff auf das *Terminologische Wörterbuch*, das in kompakter Form mit Beispielen angereicherte Definitionen grammatischer Termini liefert. Bei *ProGr@mm* sind außerdem im *Grammatischen Grundwissen* sowie im *Kontrastiv*-Modul Übungen und Kontrollaufgaben über anklickbare Fenster zu erreichen, die zum Teil einen tieferen Einblick in Sprachsystem und Sprachgebrauch ermöglichen. Manche Übungen setzen ein reflektiertes Verhältnis zu grammatischen Kategorien voraus bzw. fördern ein solches Bewusstsein.

b. Die Suche nach umfassenden und weiterführenden Kommentaren erfolgt meistens im Rahmen einer wissenschaftlichen Arbeit bzw. eines Seminars, wobei die Nutzer und Nutzerinnen aufgrund einer reflektierten Auseinandersetzung mit dem Sprachsystem und dem Sprachgebrauch vertiefte Erkenntnisse gewinnen wollen. Die Ebene der Sprachbeschreibung wird in *Grammis* (Modul *Systematische Grammatik*) und *ProGr@mm* (Modul *Grammatisches Grundwissen*) unterschiedlich vertreten. Während *Grammis* den Informationssuchenden von Anfang an mit in der Forschung divergierenden Ansichten konfrontiert und sie unter Zuhilfenahme entsprechend markierter Vertiefungstexte in die Diskussion (etwa zum Valenzbegriff oder zu Endungen finiter Verben) mit einbezieht, verfährt die eher als Lernplattform konzipierte propädeutische Grammatik *ProGr@mm* in mehreren Etappen, indem sie über die allgemeine, einfache Beschreibung der Formen und Funktionen hinaus sukzessive die Möglichkeit bietet, sich bei Bedarf genauer zu informieren, z. B. über unterschiedliche Ansichten oder über bestimmte theoretische Hintergründe. Lehrende können außerdem einen Raum für Seminare einrichten, wo Seminarleiter oder Seminarleiterin den Teilnehmerinnen und Teilnehmern themenspezifische Materialien zur Verfügung stellen kann. Das dazu gehörende *Forum* ist ein Ort des Austauschs für alle Beteiligten.

Das *Kontrastiv*-Modul geht in Bezug auf die Sprachbeschreibung noch ein Stück weiter. Es liefert zusätzliche Informationen, die vor dem Hintergrund

einer anderen Sprache von Relevanz sind, und zwar entweder weil das System des Deutschen anders gestaltet ist oder die Form-Funktion-Relation eine andere ist oder weil die grammatische Tradition in dem jeweiligen Land eine andere Herangehensweise an die Formen und Funktionen pflegt. Solche Informationen, zu denen Nutzerinnen und Nutzer nur wahlweise greifen, sind in sogenannten *Vertiefungsabsätzen* enthalten, die nach Bedarf angeklickt werden können und erst dann vollständig sichtbar werden. So zum Beispiel bei den Tempora (sowohl bei den Formen als auch bei ihrem Gebrauch) oder bei der sogenannten *Wortstellung* (die ja meistens eher eine Satzgliedstellung ist) im Zusammenhang mit der Informationsstruktur der Äußerung.

Da in *Grammis* die Themen grundsätzlich tiefgründiger behandelt werden, bleibt es für Wissenschaftler und Wissenschaftlerinnen, aber auch für fortgeschrittene Germanistinnen und Germanisten (Studierende und Lehrende) ein wichtiges, weil quasi komplettes Hilfsmittel zur Beschreibung von Formen und Strukturen des Deutschen. An die am IDS erarbeitete, dreibändige Grammatik der deutschen Sprache angelehnt, ermöglicht das darin enthaltene Modul *Systematische Grammatik* einen genauen Einblick in die grammatischen Strukturen des Deutschen aus unterschiedlichen Perspektiven. Die multimediale Aufbereitung ermöglicht eine sachgerechte Veranschaulichung der beschriebenen Formen und Strukturen, und die intuitive Hypertext-Navigation erleichtert die Zugriffswege zu verteilten Informationen. Zu erwähnen sind hier noch weitestgehend in alle Hypertexteinheiten integrierte Rubriken wie *Weiterführendes* oder *Zusätzliche Literatur in Auswahl*, die sich vor allem an Wissenschaftlerinnen und Wissenschaftler wenden, die sich vertiefend mit dem jeweiligen Thema beschäftigen wollen.

c. Die gezielte Suche nach einschlägiger Literatur erfolgt über das Modul *Grammatische Bibliographie*. Es enthält ein Recherche-System mit mehreren Suchoptionen bzw. -feldern, das eine genaue Eingabe nach verschiedenen Kriterien ermöglicht und sowohl selbstständige (Monografien, Sammelbände, Festschriften, Hochschulschriften etc.) als auch unselbstständige (Aufsätze, Artikel, Rezensionen etc.) Literatur umfasst. Abgedeckt werden gleichermaßen Print- und digitale Online-Publikationen. Die Nutzerinnen und Nutzer haben anschließend die Möglichkeit, ihre eigene Literaturliste zusammenzustellen und sie als *persönliche Merkliste* einfach zu kopieren und in eine Word-Datei einzufügen oder in standardisierten Formaten (BibTeX oder EndNote bzw. RIS) zu speichern, so dass die Daten in externen Literaturverwaltungsprogrammen gespeichert werden können. Dadurch wird das meist mühevolle Auflisten von bibliografischen Angaben bei wissenschaftlichen Arbeiten oder das Aufbereiten von Literaturlisten für Seminare erleichtert. Die Eingabe kann sich auf ein

einziges Kriterium beschränken, aber auch mehrere Kriterien kombinieren. Besonders wertvoll sind einerseits das Kriterium der untersuchten Sprachen, durch das die sprachvergleichende Dimension berücksichtigt werden kann, und andererseits das Kriterium des Objektsworts, das in einem Scrollfeld anhand einer ausführlichen Liste von Lexemen, Morphemen und Graphemen ausgewählt werden kann und eine gezielte Suche ermöglicht.

Insgesamt zeichnen sich die grammatischen Online-Angebote des IDS also durch ihre vielfältigen und differenzierten Einsatzmöglichkeiten aus, sie wenden sich an einen breiten Benutzerkreis und sind eine willkommene, attraktive Alternative, sich mit Fragen der deutschen Grammatik auf verschiedenen Ebenen zu befassen. Im Zeitalter eines mancherorts tendenziell „grammatikfeindlichen" Sprachunterrichts und einer der Linguistik gegenüber gelegentlich immer noch misstrauisch eingestellten Germanistik ist der Beitrag solcher wissenschaftlich fundierter Informationssysteme umso wichtiger. Es geht nämlich nicht nur darum, grammatische Kenntnisse zu vermitteln, sondern auch (und vor allem) zu verdeutlichen oder gar offenzulegen und das System der Sprache aus verschiedenen Perspektiven zu beleuchten. Die Navigationsoptionen, die durch das Gestaltungsprinzip „Hypertext" entstehen, leiten die Informationssuchenden viel intuitiver und gegenstandsnäher durch das Dickicht der Sprachformen als die sonst ziemlich mühsamen Umwege und labyrinthischen Irrwege über Wort- und Sachverzeichnisse traditioneller Grammatikbücher.

5 Die nächste Zukunft: Desiderata und Ausblicke

Was durch den digitalen Wandel und insbesondere durch gegenstandsgerecht eingesetzte Multimedialität im positiven Sinne möglich gemacht wird, darf nicht den Blick auf Verbesserungspotenzial und mögliche Fallstricke verstellen. Dies gilt insbesondere für Anwendungen, die auf sich im dynamischen Wandel befindliche Techniken zurückgreifen oder an der Schnittstelle zwischen sich ebenfalls in Entwicklung befindlichen akademischen (Teil-)Disziplinen liegen. Im Falle hypermedialer Online-Informationssysteme – interessanterweise werden die beiden zur Entstehungszeit von *Grammis* und *ProGr@mm* hochaktuellen Begriffe „Hypertext" und „Hypermedia" in der wissenschaftlichen Fachdiskussion mittlerweile eher selten explizit thematisiert – stammen wegweisende Impulse nicht allein aus der Linguistik bzw. Germanistik, sondern vermehrt aus Arbeiten der Nachbardisziplinen Medien- und Kommunikationswissenschaft, Computerlinguistik und Digital Humanities sowie der vielfältigen informatischen Spezialgebiete (Digitalisierung, Retrieval,

Visualisierung etc.). Nachfolgend wird deshalb auf einige interdisziplinäre Perspektiven und Desiderata in Bezug auf die grammatischen Online-Angebote des IDS eingegangen; damit einher geht eine Skizzierung ihrer inhaltlichen und funktionalen Weiterentwicklung.

5.1 Optimierung der Benutzungsoberfläche

Gerade bei Online-Anwendungen lässt sich der technisch-gestalterische Alterungsprozess kaum aufhalten. *Grammis* und *ProGr@mm* sind bereits seit Ende der 1990er Jahre im praktischen Einsatz und wurden von Anfang an regelmäßig weiterentwickelt und -gepflegt, aber das Design hat sich während der ersten zwei Jahrzehnte nur sporadisch geändert. Während die wissenschaftlichen Inhalte vom berühmt-berüchtigten Zahn der Zeit weitgehend verschont blieben, haben sich durch extreme Diversifizierung der für den Zugang nutzbaren Hardware inzwischen mannigfaltige neue Anforderungen ergeben.

An erster Stelle dürfte hier sicherlich die Verbreitung mobiler Endgeräte stehen. Zwar werden die grammatischen Online-Systeme nachweislich weiterhin zumeist am klassischen Desktop-PC genutzt, insbesondere wenn diese Nutzung im Zusammenhang mit der Abfassung umfangreicher wissenschaftlicher Studien oder der Vorbereitung einer Seminararbeit steht. Zunehmend kommen jedoch auch Laptops, Tablets oder Smartphones zum Einsatz, etwa für das punktuelle Nachschlagen unterwegs, während einer Vorlesung etc.[1] Damit verbunden sind dann zwangsläufig eine begrenztere Bildschirmgröße, reduzierte Interaktionsmechanismen (z. B. fehlende oder simulierte Tastatur, Touchscreen an Stelle eines Mauszeigers) sowie erhöhte Anforderungen an die Reaktionszeiten der Online-Systeme. Um diesen technischen bzw. medialen Entwicklungen Rechnung zu tragen, wurde der Internetauftritt der grammatischen IDS-Angebote – auch unter Einbeziehung von Erkenntnissen, die von *Grammis* inspirierte Portale für andere Sprachen gesammelt haben (vgl. z. B. Landsbergen, Tiberius & Dernison 2014) – ab 2018 konsequent überarbeitet und responsiv gestaltet. Beim Aufruf einer Seite wird seither automatisch erkannt, welcher Art das benutzte Endgerät ist, und die Anzeige – durch Änderung der Schriftgrößen, Skalierung und Positionierung von Grafiken,

1 Entsprechend detaillierte Aufschlüsselungen der *Grammis*-Onlinezugriffe nach Endgerätetypen und Betriebssystemen – und darüber hinaus nach diversen anderen relevanten Metadaten – bildeten den Ausgangspunkt der beschriebenen Optimierungen sowie daran anschließender Überlegungen zur Konzeption einer speziell für den mobilen Einsatz optimierten Smartphone-App.

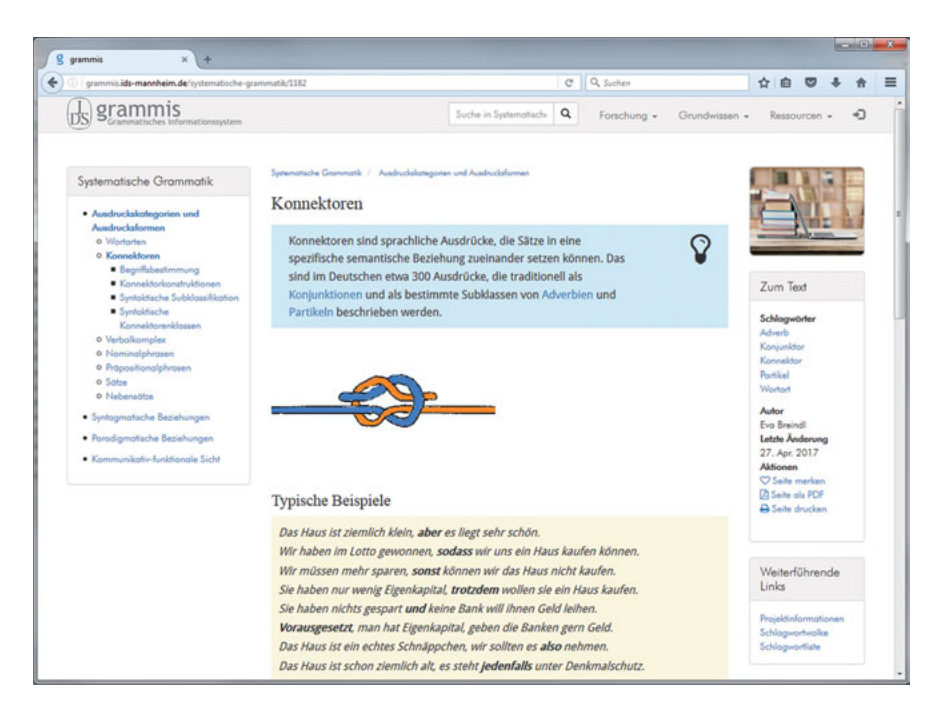


Abb. 12.1: Responsives Design und funktional orientierte Gliederung.

Ein- bzw. Ausblenden von Bedienelementen wie Menüleisten etc. – dynamisch angepasst und optimiert (vgl. Abb. 12.1).

Im Gegenzug hat sich abseits des mobilen Einsatzbereichs, also bei der stationären Nutzung am Schreibtisch, die Qualität der eingesetzten Anzeigegeräte in jüngerer Zeit tendenziell erhöht; dies betrifft Parameter wie Bildschirmgröße, Auflösung oder Farbtiefe. Ein darauf reagierendes Layout mit vielfältigeren Differenzierungs- und Hervorhebungsmöglichkeiten erscheint deshalb nicht nur möglich, sondern aus didaktischer Perspektive geradezu willkommen. Auch die optimale Nutzung der maximalen Bildschirmgröße für ein funktional sinnvolles Nebeneinander von Primärinhalten (im Browser-Hauptfenster vs. in temporär eingeblendeten Modalfenstern), Metadaten (Autor, Änderungsdatum, Zitation, Verschlagwortung, Weiterführendes etc.) und Navigationshilfen (hypertextuelle Verlinkungen, hierarchische Übersichten, Objekt- und Schlagwortlisten, Volltext- und Schlagwortsuchen) galt es auszuloten. Dabei wurde die grundsätzlich bewährte Ausrichtung – Beschränkung auf das Wesentliche, gutes Webdesign ist einfaches Webdesign – kombiniert mit der Öffnung für Neuerungen, insbesondere bei der Visualisierung von Strukturen und Querverweisen.

Neben visuell beförderten stehen auch sprachtechnologisch innovative Benutzungszugänge auf der Agenda der grammatischen IDS-Angebote. Eine zukünftig zentrale Strategie besteht dabei in der sogenannten „beispielbasierten Abfrage“ (*example-based querying*). Dieses von Augustinus, Vandeghinste & Vanallemeersch (2016) für annotierte Baumbanken erprobte Suchverfahren erlaubt den gezielten Zugriff auf linguistisch klassifizierte Informationseinheiten ohne zeitintensives Ausfüllen spezialisierter Suchformulare. Darüber hinaus kann es in Situationen zum Einsatz kommen, in denen Anwender und Anwenderinnen aufgrund terminologischer Unsicherheiten das grammatische Phänomen, nach dem sie suchen, nicht zweifelsfrei benennen können. Ein gleichermaßen regelbasierter wie maschinell lernfähiger Algorithmus analysiert in diesen Fällen vom Anwender eingegebene Beispiele aus der Alltagssprache (Beispiel: „Anfang diesen Jahres oder Ende dieses Jahres?“) und liefert – unter Rückgriff auf eine morphosyntaktisch annotierte Belegdatenbank sowie eine feingranulare terminologische Systematik (beide werden weiter unten noch angesprochen) – Informationseinheiten mit syntaktisch ähnlichen Beispielsätzen bzw. dazu einschlägigen Schlagwörtern (zum genannten Beispiel etwa „Genitiv“ und „Demonstrativ-Artikel“).

5.2 Inhaltliche Integration und Erweiterung

Popularität und Akzeptanz der grammatischen Online-Systeme führten über die Jahre zu einem steten Wachstum der in *Grammis* und *ProGr@mm* aufgenommenen thematischen Module. Hinzu kamen eigenständige Portale, die phänomenspezifische Forschungsdaten und deren wissenschaftliche Interpretation bündeln, beispielsweise in Form des digitalen Valenzwörterbuchs *E-VALBU* (Schneider 2008) oder der Genitivdatenbank *GenitivDB* (Schneider 2014; Hansen-Morath, Konopka & Schneider 2016). Diese Verteilung auf mehrere Plattformen erschwerte zunehmend das gezielte Auffinden der insgesamt für eine Fragestellung relevanten Informationseinheiten, ebenso wie die informatische Pflege der zugrunde liegenden Autorenumgebungen.

Seit 2018 bündelt das neue *Grammis* sämtliche grammatischen Online-Inhalte des IDS unter einer einheitlichen, übersichtlichen Oberfläche (vgl. Abb. 12.2). Wesentlich ist eine inhaltlich und klassifikatorisch motivierte Dreiteilung in folgende Hauptbereiche:

- Unter der Rubrik **Forschung** finden sich die Ergebnisse aktueller und abgeschlossener grammatischer Forschungsprojekte des Instituts für Deutsche Sprache, die sich explizit an ein linguistisch ausgebildetes Fachpublikum wenden. Hierzu zählen die ehemals zentrale *Systematische Grammatik*, die bereits erwähnte *Korpusgestützte Grammatik*, die sich der

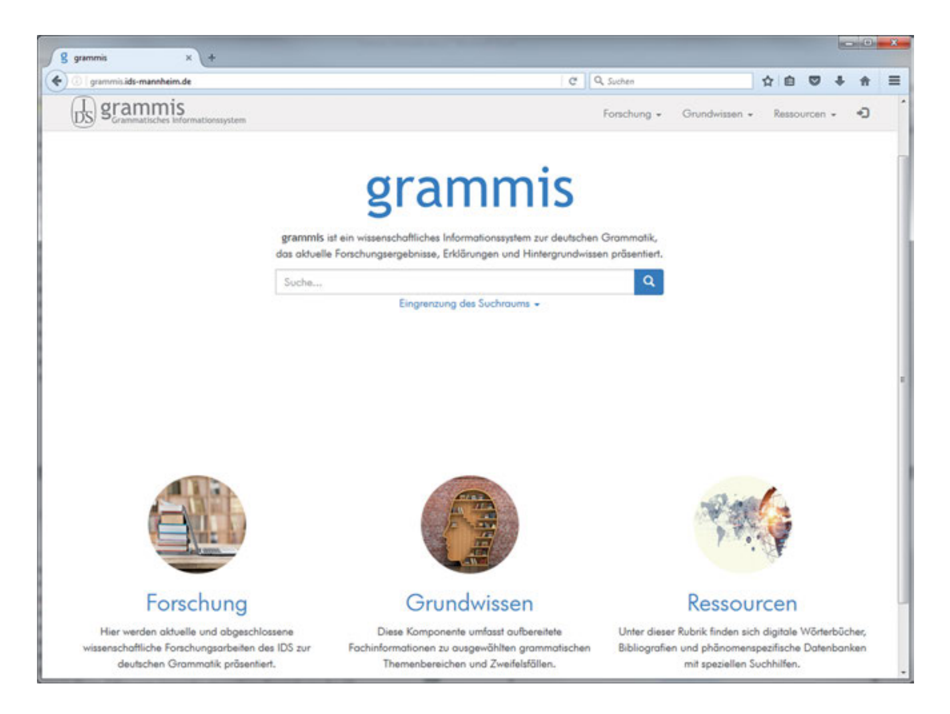


Abb. 12.2: Einstiegsseite und Hauptbereiche.

Erforschung grammatischer Variation im standardsprachlichen und standardnahen Deutsch widmet, sowie neuere Module etwa zu den Bereichen *Kontrastive Grammatik* und *Wortphonologie*. Abgerundet wird die Rubrik von einer onomasiologisch aufgebauten grammatischen Terminologiekomponente.

- Die Rubrik **Grundwissen** versammelt die ursprünglich in *ProGr@mm* enthaltenen didaktisch aufbereiteten Module, namentlich das *Grammatische Grundwissen* sowie die im Projekt EuroGr@mm erarbeitete *Kontrastive Sicht*. Darüber hinaus fallen in diese Rubrik die ebenfalls an eine breitere Öffentlichkeit gerichtete *Grammatik in Fragen und Antworten* sowie das offizielle Regelwerk zur deutschen Rechtschreibung.
- Die Rubrik **Ressourcen** schließlich integriert digitale Einzelwörterbücher, thematisch motivierte Bibliografien und phänomenspezifische Datenbanken mit speziellen Suchhilfen (etwa *GenitivDB* für die Genitivforschung oder ein mit morphologischen und phonologischen Metadaten angereicherter Grundwortschatz für den Rechtschreibunterricht).

Sämtliche Rubriken werden kontinuierlich gepflegt und ausgebaut. Dabei ergeben sich immer wieder Situationen, in denen es nicht mit einer reinen Sub-

summierung existierender digitaler Inhalte unter einer neuen einheitlichen Oberfläche getan ist, sondern in denen inhaltliche Arbeit notwendig wird. Als Beispiel sei die Zusammenführung der über die Jahre gewachsenen Konnektoren-Ressourcen genannt: Das ehemals eigenständige *Wörterbuch der Konnektoren* lieferte grammatisch relevante Informationen zu Konnektoren wie *aber*, *weil*, *wohlgemerkt*, *sogar* oder *geschweige denn*, basierend auf den Ergebnissen des ersten Teilprojekts des *Handbuchs der deutschen Konnektoren* (Pasch et al. 2003; HDK-1). Die ehemals in die *Systematische Grammatik* integrierte *Datenbank der deutschen Konnektoren* versammelte Informationen zu allen Konnektoren, die im Registerteil von des zweiten Teilprojekts HDK-2 (Breindl, Volodina & Waßner 2014) aufgelistet sind. Diese galt es inhaltlich abzugleichen und in einer einheitlichen Systematik zu verorten. Darüber hinaus wurden Beispiele und korpusbasierte Belege ergänzt, so dass nunmehr alle syntaktischen und semantischen Varianten umfassend illustriert sind.

5.3 Terminologische Aspekte

Bereits die frühesten *Grammis*-Implementierungen hatten mit terminologischer Varianz umzugehen. Einerseits gegründet auf den Arbeiten und Festlegungen der *Grammatik der deutschen Sprache* (Zifonun et al. 1997), verfolgte das Autorenteam andererseits von Anfang an das Ziel, in die hypermediale Überarbeitung auch neuere Erkenntnisse und methodische Weiterentwicklungen einfließen zu lassen. Spätestens mit der expliziten Öffnung für ein breiteres sprachinteressiertes Nutzerspektrum kommt der Umstand hinzu, dass nicht nur den einzelnen *Grammis-/ProGr@mm*-Modulen unterschiedliche theoretische Klassifizierungen zugrunde liegen, sondern insbesondere auf Leserseite mit einem uneinheitlichen Vorwissen – sprich: Terminologieinventar – gerechnet werden muss. Die internationale (Teil-)Ausrichtung sowie eine fortlaufende Einbeziehung zusätzlicher Schwerpunktthemen und wissenschaftlicher Autorinnen und Autoren verstärken diesen Aspekt kontinuierlich.

Zielsetzung des grammatischen Informationssystems bleibt die innovative Präsentation potenziell heterogener multimedialer Sprachressourcen und Forschungsergebnisse. Dabei gilt es, ein breites Nutzerspektrum vom interessierten Laien über professionell mit Sprache befasste Berufsgruppen bis hin zu Sprachwissenschaftlern und -wissenschaftlerinnen unterschiedlicher theoretischer Ausrichtung abzudecken. Vor diesem Hintergrund lautet die Kernfrage: Wie findet die oder der Informationssuchende – im Einzelfall nicht nur mit unterschiedlich stark ausgeprägter linguistischer Bildung, sondern insbesondere mit theorieabhängig variablem terminologischem Vokabular – die für sie oder ihn passende Stecknadel im Heuhaufen fachwissenschaftlicher Inhalte?

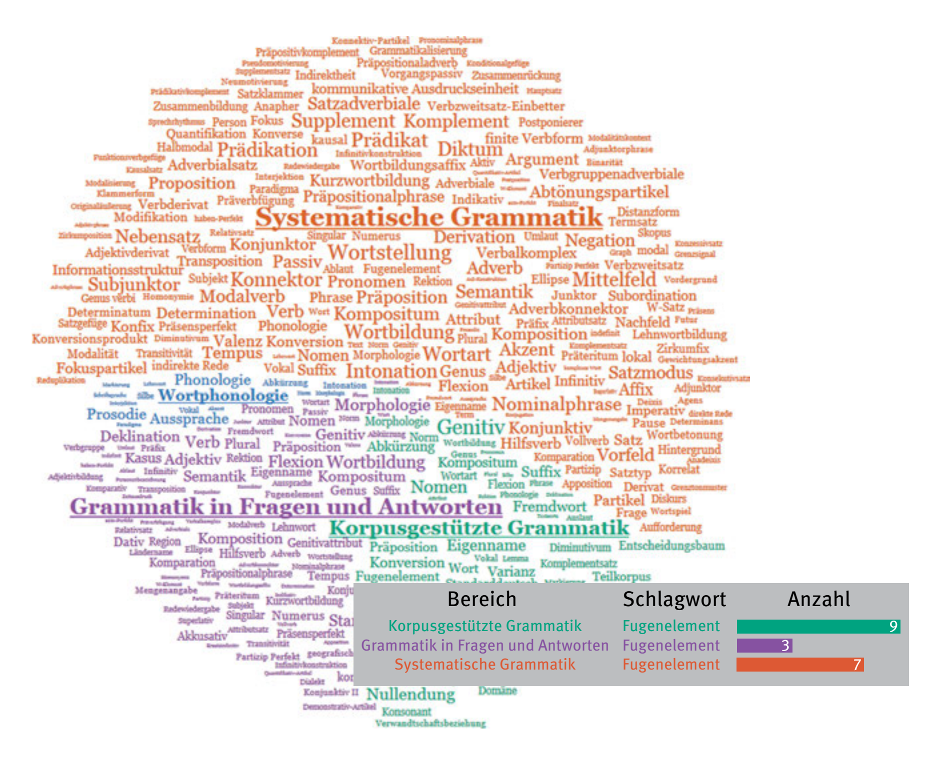


Abb. 12.3: Schlagwortwolke als modulübergreifende Navigationshilfe.

Eine zentrale Bedeutung im gegenwärtigen wie zukünftigen *Grammis*-Portal kommt deshalb einer konsistenten, terminologieübergreifenden Vernetzung von inhaltlichen Modulen zu, die aus grammatiktheoretischer Perspektive nicht durchgehend und in allen Einzelheiten einheitlich ausgerichtet sein können. Um eine automatisierbare Bezugnahme zwischen mit unterschiedlichem terminologischem Inventar formulierten, aber das gleiche sprachliche Phänomen beschreibenden Inhalten zu befördern, bildet eine onomasiologisch konzipierte Terminologiedatenbank das Rückgrat der *Grammis*-Recherche (Sejane 2010; Suchowolec, Lang & Schneider 2016; Suchowolec et al. 2017). Unter Rückgriff auf aus Ontologien und Thesauri bekannten Äquivalenz-, Assoziations- und hierarchischen Relationen vermittelt diese in terminologischen Zweifelsfällen zwischen System und Nutzer. Neben einer konsistenten Makrostruktur (Vernetzung von Begriffen) des Fachvokabulars enthält die Terminologiedatenbank auch eine Vielzahl von erklärenden Kurzartikeln zu relevanten grammatischen Termini in einer normierten Mikrostruktur (Kurzdefinition, Erläuterungstext, Bestand, Beispiele/Korpusbelege, weiterführende Hinweise) und integriert bzw. erweitert zu diesem Zweck das in Abschnitt 4 beschriebene *Terminologische*

Wörterbuch. Außerdem definiert sie den Gesamtbestand aller für die Verschlagwortung von *Grammis*-Informationseinheiten einsetzbaren Termini.

Abbildung 12.3 illustriert einen Anwendungsfall dieser (manuellen oder maschinell unterstützten) Verschlagwortung für die grafisch unterstützte Navigation. Eine modulübergreifende Schlagwortwolke – ein Hybrid aus *Tag Cloud* und *Pie Chart* – visualisiert die Verteilung aller textspezifischen Schlagwörter über die einzelnen *Grammis*-Module. Dabei korrespondieren die Schriftfarben mit den Modulen und die Schriftgröße mit der Häufigkeit. Per Mausklick lassen sich anschließend die mit dem jeweiligen Schlagwort verknüpften Informationseinheiten aufrufen. Alternativ steht eine alphabetisch geordnete Schlagwortliste zur Verfügung.

5.4 Qualitätssicherung für Sprachbelege

Der Rückgriff auf authentische Sprachbelege ist für die linguistische Forschung wie für die Vermittlung von Forschungsergebnissen wichtig und bildet einen entsprechend wesentlichen Bestandteil der grammatischen Informationssysteme. Allerdings kann er im Rahmen der Vermittlung bzw. Illustration zu Fehlgriffen führen: Informationssuchende stoßen erfahrungsgemäß immer wieder auf Beispiele, die sie nicht auf Anhieb verstehen, und zwar entweder wegen ihrer Länge oder Komplexität – verglichen mit der zu illustrierenden Form/Funktion – oder weil sie zu sehr kontextuell angebunden sind und ein Welt- und Fachwissen voraussetzen, das nicht jeder Nutzer oder jede Nutzerin haben kann. Dieses Problem besteht unabhängig von der sonstigen Rolle des Korpus: Ob es darüber hinaus als empirische Datenbasis für statistische Analysen dient oder „nur" eine Kontroll-Funktion haben soll – die ausgewählten Belege, die in der Online-Darstellung die grammatischen Strukturen illustrieren sollen, müssen relativ einfach und gleich einleuchtend, d. h. leicht zu interpretieren sein.

Diesem Umstand wird sich in *Grammis* zukünftig eine spezielle Belegdatenbank widmen. Konzipiert sowohl aus der Forschersicht (die in den meisten Fällen authentisches Sprachmaterial variabler Komplexität präferiert, um den Gegenstandsbereich angemessen beurteilen zu können) als auch aus der Perspektive von Sprachlernenden (die in erster Linie an didaktisch sinnvollen Beispielen interessiert sind, welche im Einzelfall auch konstruiert sein dürfen), integriert sie beide genannten Belegtypen. Einen wesentlichen Grundstock bezieht sie dabei aus den in *Grammis* bereits annotierten Beispielsätzen. Darüber hinaus fließen kontinuierlich Belege ein, die im Rahmen flankierender Projekte und Studien zu speziellen Fragestellungen erarbeitet werden.

Die Belegdatenbank kombiniert authentische und konstruierte Beispiele mit diversen Metadaten: Neben morphosyntaktischen und ggf. weiteren Annotationen zählen hierzu in erster Linie eine detaillierte linguistisch motivierte Verschlagwortung sowie didaktisch aussagekräftige Klassifizierungen. Im Ergebnis soll die Datenbank die *Grammis*-Autoren und -Autorinnen bei der Auswahl stimmiger Belege unterstützen und für die *Grammis*-Nutzerinnen und -Nutzern die Navigation zu inhaltlich passenden Informationseinheiten befördern, etwa vermittels der weiter oben angesprochenen „beispielbasierten Abfrage".

5.5 Vermittlung in den Kontrastsprachen

Abschließend soll noch ein Aspekt erwähnt werden, der zwar kritisch angegangen wird, aber dennoch grundsätzlich ohne weiteres verbessert werden kann: Es handelt sich hier um die Erweiterung und „Fruchtbarmachung" der aus der kontrastiven Perspektive geleisteten Arbeit. Unsere Erfahrung zeigt, dass die Lernenden bzw. Studierenden aus der jeweiligen Muttersprache Schwierigkeiten haben, die Module auf Deutsch zu verstehen. Dies hängt zum Teil mit der entsprechenden Lehrtradition zusammen: Wenn Grammatik nur in der Muttersprache (d. h. nicht auf Deutsch) gelehrt wird, wird der Gebrauch von Deutsch als Metasprache als zusätzliches Hindernis empfunden. Dies schränkt den Einsatz der Propädeutischen Grammatik *ProGr@mm* im Grundstudium ein sowie auch den Verweis auf andere *Grammis*-Module bei Fortgeschrittenen. Deshalb hier zum Abschluss die Gretchen-Frage: Wäre es nicht relevant und wünschenswert, auf Basis des im *Kontrastiv*-Modul erarbeiteten Modells kontrastive – oder zumindest kontrastiv angelegte – Grammatiken des Deutschen in den jeweiligen Kontrastsprachen zu erstellen?

Wie dem auch sei: Der Blick von außen auf die deutsche Sprache, der Bedarf an einer umfassenden, erklärenden Beschreibung der grammatischen Strukturen des Deutschen, die Wirkung des IDS nach außen und nicht zuletzt der digitale Wandel in den Wissenschaften sind wichtige Pfeiler der hier vorgestellten Online-Angebote. Ihre Erhaltung, Umgestaltung und Weiterentwicklung sind eine anspruchsvolle Aufgabe, die – nicht nur, aber eben auch aus Perspektive der Germanistik im Ausland – ernst genommen werden muss!

Literatur

Augustin, Hagen (2009): EuroGr@mm – Internetprojekt und europäisches Forschungsnetzwerk. *Sprachreport* 1, 24–27.

Augustinus, Liesbeth, Vincent Vandeghinste & Tom Vanallemeersch (2016): Poly-gretel: Crosslingual example-based querying of syntactic constructions. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association*, 3549–3554.

Breindl, Eva, Anna Volodina & Ulrich Hermann Waßner (2014): *Handbuch der deutschen Konnektoren 2. Semantik der deutschen Satzverknüpfer.* Berlin, Boston: de Gruyter.

Breindl, Eva, Roman Schneider & Bruno Strecker (2000): GRAMMIS – ein Projekt stellt sich vor. *Sprachreport. Informationen und Meinungen zur deutschen Sprache* 1, 19–24.

Bubenhofer, Noah, Marek Konopka & Roman Schneider (2014): *Präliminarien einer Korpusgrammatik. Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache (CLIP).* Tübingen: Narr.

Dalmas, Martine, Cathrine Fabricius-Hansen & Horst Schwinn (Hrsg.) (2016): *Variation im europäischen Kontrast. Untersuchungen zum Satzanfang im Deutschen, Französischen, Norwegischen, Polnischen und Ungarischen*. Berlin, Boston: de Gruyter. (= Konvergenz und Divergenz 5).

Hansen-Morath, Sandra, Marek Konopka & Roman Schneider (2016): Empirische Analysen zur Genitivvariation mit GenitivDB 2.0. In: Jianhua Zhu, Jin Zhao & Michael Szurawitzki, (Hrsg.), *Akten des XIII. Internationalen Germanistenkongresses Shanghai 2015: Germanistik zwischen Tradition und Innovation, Band 2*, Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Wien: Peter Lang.

Konopka, Marek & Roman Schneider (Hrsg.) (2012): *Grammatische Stolpersteine digital. Festschrift für Bruno Strecker zum 65. Geburtstag.* Mannheim: Institut für Deutsche Sprache (IDS).

Konopka, Marek (2006): Grammatik in Fragen und Antworten. *Sprachreport. Informationen und Meinungen zur deutschen Sprache* 3, 9–12.

Landsbergen, Frank, Carole Tiberius & Roderik Dernison (2014): Taalportaal: An online grammar of Dutch and Frisian. In: Nicoletta Calzolari et al. (Hrsg.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 26–31. Reykjavik: European Language Resources Association (ELRA).

Pasch, Renate, Ursula Brauße, Eva Breindl & Ulrich Hermann Waßner (2003): *Handbuch der deutschen Konnektoren. Linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfer (Konjunktionen, Satzadverbien und Partikeln).* Berlin/New York: de Gruyter.

Schneider, Roman (2004): *Benutzeradaptive Systeme im Internet: Informieren und Lernen mit GRAMMIS und ProGr@mm.* Mannheim: IDS. (= amades 4/04).

Schneider, Roman (2008): E-VALBU: Advanced SQL/XML processing of dictionary data using an object-relational XML database. *Sprache und Datenverarbeitung, International Journal for Language Data Processing*, 33–44.

Schneider, Roman (2014): *GenitivDB — a corpus-generated database for German genitive classification. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), European Language Resources Association.*

Schneider, Roman & Horst Schwinn (2014): Hypertext, Wissensnetz und Datenbank: Die Web-Informationssysteme grammis und ProGr@mm. In: Institut für Deutsche Sprache (Hrsg.), *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, 337–346. Mannheim: IDS.

Schwinn, Horst (2005): ProGr@mm. Die propädeutische Grammatik der deutschen Sprache. *Sprachreport* 3, 17–19.

Sejane, Ineta (2010): *Wissensrepräsentation Linguistik. Modellierung, Potenzial und Grenzen am Beispiel der Ontologie zur deutschen Grammatik im GRAMMIS-Informationssystem des IDS Mannheim*. Heidelberg: Universitätsbibliothek.

Storrer, Angelika (1997): Vom Text zum Hypertext. Die Produktion von Hypertexten auf der Basis traditioneller wissenschaftlicher Texte. In: Dagmar Knorr & Eva-Maria Jacobs (Hrsg.), *Textproduktion in elektronischen Umgebungen*, 121–139. Frankfurt a. M.: Lang.

Strecker, Bruno (1998): Hypertext: Chance oder Herausforderung für die Grammatikschreibung? In: Angelika Storrer & Bettina Harriehausen (Hrsg.), *Hypermedia für Lexikon und Grammatik*, 21–28. Tübingen: Narr. (= Studien zur deutschen Sprache).

Suchowolec, Karolina, Christian Lang & Roman Schneider (2016): Re-designing online terminology resources for German grammar. In: Philipp Mayr, Douglas Tudhope, Koraljka Golub, Christian Wartena, & Ernesto William De Luca (Hrsg.), *Proceedings of the 15th European Networked Knowledge Organization Systems Workshop* (NKOS 2016), 59–63. http://ceur-ws.org/Vol-1676/

Suchowolec, Karolina, Christian Lang, Roman Schneider & Horst. Schwinn (2017): Shifting complexity from text to data model. Adding machine-oriented features to a human-oriented terminology resource. In: J. Gracia, F. Bond, J. P. McCrae, P. Buitelaar, C. Chiarcos & S. Hellmann (Hrsg.), *Language, data, and knowledge. Lecture notes in artificial intelligence*, 203–212. Berlin, Heidelberg, New York: Springer. doi:10.1007/978-3-319-59888-8.

Zifonun, Gisela, Ludger Hoffmann, Bruno Strecker, Joachim Ballweg, Ursula Brauße, Eva Breindl, Ulrich Engel, Helmut Frosch, Ursula Hoberg & Klaus Vorderwülbecke (1997): *Grammatik der deutschen Sprache*. Berlin: de Gruyter.

IV Annotation und Modellierung

C. M. Sperberg-McQueen

13 Kernideen der deskriptiven Textauszeichnung

Abstract: Die weitverbreiteten Datenformate SGML und XML wurden als Träger für deskriptive Textauszeichnung entwickelt. XML bietet eine sehr einfache Syntax, eine flexible Wahl an Datenmodellen, sowie die Möglichkeit, ein eigenes Auszeichnungssystem zu definieren und mechanisch nach Fehlern zu überprüfen. Die Freiheit, selbst das Datenformat zu bestimmen, gibt dem Anwender von XML sowohl die Möglichkeit, Textphänomene mit Auszeichnungen festzuhalten, die keine kommerzielle Software vordefiniert, als auch eine Eigenverantwortung für die eigenen Daten. Beides ist für die wissenschaftliche Arbeit mit Texten unerlässlich.

Keywords: Annotation, Deskriptive Textauszeichnung, Sprachkorpora, XML

1 Einleitung

Vor etwa dreißig Jahren (am 15. Oktober 1986) wurde von der internationalen Organisation der Standardisierung ISO eine Norm verabschiedet, mit Namen *ISO 8879: Information processing – Text and office systems – Standard Generalized Markup Language* (SGML), zu Deutsch Informationsverarbeitung – Bürosysteme – Standardauszeichungssprache (SGML).

SGML will eine technische Basis für die deskriptive Textauszeichnung (*descriptive markup*) liefern; es lohnt sich vielleicht nach dreißig Jahren ein Versuch, einige der Kernideen dieses Begriffs zu erläutern und zu würdigen. Diese Kernideen helfen, die Zählebigkeit des Begriffs und der ISO-Norm zu erklären.

Wie der Titel der ISO-Norm andeutet, stufte ISO die SGML als Beitrag zur Automatisierung der Büroarbeit ein. Aber schon in der Arbeitsgruppe, die diese Norm entwickelte, hatte man einen weiteren Einsatz anvisiert, vor allem im Verlagswesen und im Druckwesen, vorzugsweise in der technischen Redaktion

C. M. Sperberg-McQueen, Black Mesa Technologies, 259 State Road 399, Española NM 87532, E-Mail: cmsmcq@blackmesatech.com

https://doi.org/10.1515/9783110538663-014

(*technical publishing*), die nicht Bücher oder Zeitschriften, sondern Gebrauchsanleitungen, Produkthandbücher usw. erstellt.

Heute werden aber SGML und ihre Nachfolgerin XML (*Extensible Markup Language*, 1996 vom World Wide Web Consortium (W3C) verabschiedet) sehr viel breiter verwendet. In Verlagen und in der technischen Redaktion werden immer noch SGML und XML angewandt, sie werden aber auch für Grafik (*Scalable Vector Graphics* (SVG) ebenfalls vom W3C definiert) und Bibliotheksdaten (bei der Digitalisierung von Texten und auch beim Austausch von Metadaten) gebraucht. Datenbanksysteme setzen XML für Datenaustauch ein und haben auch inzwischen XML-orientierte Abfragesprachen, entweder eine Erweiterung zu SQL namens SQL-XML oder eine eigene XML-orientierte Sprache, XQuery (W3C). Sogar bei Tankstellen wendet man XML an: Die Tanksäule und die Kasse senden sich gegenseitig Messages, die nicht immer aber meistens (so wird von Fachleuten berichtet) mit XML kodiert sind. In der Linguistik hat sich die strukturierte Auszeichnung mit SGML oder XML zum de facto-Standard für die Verarbeitung natürlichsprachlicher Ressourcen entwickelt. Große Textkorpora wie DeReKo werden in einem XML-Format (s. Lüngen & Sperberg-McQueen 2012) verwaltet, multi-dimensionale Annotationen (morphosyntaktisch, semantisch etc.) werden durch Tagger in XML erstellt, selbst Schnittstellen zwischen sprachverarbeitenden Programmen werden oft mit XML spezifiziert. XML dient als Basis aktueller empirischer und sprachtechnologischer Arbeiten.

XML ist also praktisch das geworden, was man schon 1996 schlagwortartig vorausgesagt hat: ein ASCII für das 21. Jahrhundert. Sie findet fast überall Verwendung, wo man überhaupt Informationstechnologie anwendet. Eine breite, vollständige Bestandsaufnahme der Anwendungen von SGML, XML oder allgemeiner von der deskriptiven Textauszeichnung ist wegen der Breite der Anwendungsgebiete hier nicht möglich. Ich beschränke mich darauf, einige Kernideen von SGML und XML zu identifizieren.

2 Deskriptives Markup

SGML hat drei Wurzeln, denen ich nachgehen möchte: in der Büroautomatisierung, in der technischen Redaktion und im Lichtsatz.

Die Vorläufersprache GML (*Generalized Markup Language*) entstand einem IBM-Projekt, in dem man Software für Rechtskanzleien entwickeln wollte. Um die in der Kanzlei verfertigen Dokumente zu verwalten, sollte diese Software die Möglichkeit bieten, Texte zu verarbeiten und Dokumente auszudrucken. Die Dokumente sollten aber auch archiviert und durchsuchbar gemacht wer-

den, damit man (etwa bei Änderung der Rechtslage) bequem ausfindig mache könnte, welche Auftraggeber von der Änderung betroffen wären. Textverarbeitungssysteme und Wortprozessors existierten schon, wie auch Volltextdatenbanken. Aber die Datenformate, die man für die einen einsetzte, hatten mit den Datenformaten der anderen nichts gemeinsam. GML ist aus dem Bestreben entstanden, diese zwei Anwendungen auf die gleichen Daten zugreifen und die gleichen Daten verarbeiten zu lassen. In der Büroautomatisierung suchte man also eine anwendungsunabhängige oder wiederverwendbare Dateistruktur.

In der technischen Redaktion verwaltet man oft eine sehr umfangreiche Produktdokumentation mit tausenden oder hunderttausenden von Seiten. Aus diesem Material müssen Dokumente oft in verschiedenen Versionen erzeugt werden, je nach Modell des Produkts, wobei eine Konsistenzüberprüfung der Ausgabe nicht unterbleiben darf. Viele Dokumente, wie z. B. Computerhandbücher, wollen auch mehrfach gestaltet werden: einmal auf Papier (das waren noch Zeiten!), aber auch auf Bildschirm. Heutzutage kann man davon ausgehen, dass alles, was man auf Papier darstellen kann, praktisch sehr ähnlich auf dem Bildschirm zu bringen ist. In den 1980er Jahren war das aber nicht so. Es werden auch ständig neue Versionen der Dokumente verlangt, oft mit Textänderungen, aber oft auch ohne jede Änderung des Wortlauts, wenn etwa die Firma allen Dokumenten eine neue graphische Gestaltung (ein neues Look) geben will. Es kann sein, dass z. B. die Größe, die Schriftart, oder die Positionierung der Kapitelüberschriften anders gestaltet wird. Wollte man eine solche Änderung manuell durchführen, so müsste man jedes Dokument im Editor aufmachen, alle Kapitelüberschriften finden und ändern, und das Dokument wieder speichern. Die Erfahrung hat gelehrt, dass ein solcher Versuch nicht gut endet. Man überspringt das eine oder das andere Dokument; man findet nicht alle Kapitelüberschriften, man führt aus Versehen eine falsche neue Schriftart ein oder eine globale Änderung trifft auch andere Stellen des Texts, die keine Kapitelüberschriften sind und diese neue Gestaltung nicht haben sollen. Manuell gelingen solche Massenänderungen der Textgestaltung nicht; man muss sie automatisieren können. Das lässt sich am einfachsten bewerkstelligen, wenn die Gestaltungsanweisungen für Kapitelüberschriften nicht im Dokument mit den Kapitelüberschriften angegeben werden (was auch mehrfach redundant ist), sondern nur einmal extern spezifiziert werden. In der technischen Redaktion suchte man also die Möglichkeit, Text und Verarbeitungsanweisungen getrennt zu speichern, damit man Letztere unabhängig von dem Ersten ändern kann.

Die Lichtsatzdienstleister hatten ein scheinbar anderes Problem. Jede Lichtsatzmaschine hatte derzeit eigene Anweisungskodes, meist sehr hardwarespezifisch, sehr hardwarenah. Beim Umgang mit solchen Kodes entstehen

leicht Kodierungsfehler. Dass jede Maschine eigene Anweisungen hatte, führte zu einem fragmentierten Markt. Dass die Kodierungsfehler so leicht waren, führte zu unerwünschten Ausgaben, denn Kodierungsfehler sind bei solchen Maschinen manchmal sehr teuer. Die Dienstleister wollten ihre Kundenbasis erweitern; die Kunden wollten eine größere Auswahl an Anbietern haben. Beide Seiten hatten also ein gemeinsames Interesse an einem größeren Markt. Und beide hatten ein gemeinsames Interesse daran, Fehler in der Kodierung so früh wie möglich zu entdecken und beseitigen zu können. Im Lichtsatzsektor suchte man also eine Art Hardwareunabhängigkeit und zu gleicher Zeit eine Art Lieferantenunabhängigkeit der Daten.

Man hat sich bald darauf geeinigt, dass für diese drei Probleme eine gemeinsame Lösung zu finden war, und zwar in einem Dateiformat, nicht etwa in einer Schnittstelle zwischen Programmen. Der Vorsatz, die Anwendungsunabhängigkeit, die Wiederverwendbarkeit und die Hardwareunabhängigkeit der Daten durch ein Datenformat und nicht durch eine Schnittstelle zu gewährleisten, unterscheidet SGML grundsätzlich von der zeitgenössischen Datenbanktheorie, wo man bei der Entwicklung von SQL gerade den entgegengesetzten Weg gegangen ist.

Die aus diesen drei Wurzeln entsprossene Lösung heißt generisches Markup oder deskriptive Textauszeichnung. Mit Markup oder Textauszeichnung meint man Metadaten, die man zur Beschreibung der anvisierten Verarbeitung in einen Text einfügt. Markup in diesem allgemeinen Sinn existierte schon im Wiegendruckzeitalter, wo man am Rande einer Handschrift einfache Setzeranweisungen eingefügt hat. Im Laufe der Zeit wurde die Druckseitengestaltung komplizierter und das Markup routinierter. Man gibt Schriftart, Schriftgröße und Satzbreite vor, und man bemüht sich, für Schriftarten mehr oder weniger standardisierte Namen einzuführen, wie man auch für Größen- und Breitenanweisungen standardisierte Maßsysteme entwickelt. Wenn man die Anweisung „Palatino, 10 Punkt, 11 Punkt Durchschuss, Satzspiegel 30 em“ gibt, so hat man die Seite mehr oder weniger vollständig beschrieben. Die ersten Formatierungsprogramme haben für diese und andere Eigenschaften der Seite eigene Kommandos bereitgestellt.

Deskriptives Markup nennt man Markup, das die Daten beschreibt, anstatt einem Programm Anweisungen zu geben. Typischerweise ist deskriptives Markup generisch, nicht gerätespezifisch oder anwendungsspezifisch; deklarativ, statt imperativ zu sein; logisch und nicht formatierungsorientiert. Man orientiert sich an dem, was man mehr oder weniger naiv „die logische Struktur eines Texts“ nennt. (Naiv deswegen, weil nicht ohne Weiteres anzunehmen ist, dass es nur eine logische Struktur gibt, und nicht mehrere.)

In den 1970er Jahren bildete eine Druckindustriegruppe in den USA namens *Graphic Communications Association* (GCA) eine Arbeitsgruppe namens *GenCode*

Committee. Zur Aufgabe hatte diese Arbeitsgruppe die Formulierung eines generischen Kodes für den Satz. Das Generische an diesem Kode wäre, dass er allgemein anwendbare Kommandos definieren sollte, wie etwa „Neue Seite!" oder „Neue Schriftart!", die man einsetzen könnte, statt irgendeine hardwarespezifische Escape-Sequenz wie „ESC x 234 y" schreiben zu müssen. Schon an diesem Beispiel wird klar, dass eine generische Kodierung der Typographie leichter zu verstehen sein kann, als die maschineneigene Kodierung.

Diese Arbeitsgruppe von GCA ist bald zu dem Ergebnis gekommen, dass ihr Auftrag sinnlos sei, und dass das Grundproblem anders zu lösen sei. Sie hat sich geweigert, das ihr aufgetragene Problem auf die ihr vorgeschriebene Weise zu lösen, und hat stattdessen ein anderes und wichtigeres Problem gelöst. Einen einheitlichen Kode, der für alle und alles funktioniert, gebe es nicht und könne es nicht geben. Der Versuch, einen solchen Kode zu schaffen wäre also vergebliche Mühe. Jeder vorstellbare Kode reicht vielleicht für die eine Anwendung, ist aber für eine andere nicht gut geeignet oder umgekehrt. Ein Kode, der für alle Interessierten brauchbar ist, könne man aber nicht erreichen.

Ihre Lösung könnte man eine Flucht in die Metaebene nennen. Statt selbst einen Kode, eine Kommandosprache für den Satz zu definieren, hat die Arbeitsgruppe einen Metakode, eine Metasprache definiert: eine Sprache zur Definierung anderer Sprachen. Mit dieser Metasprache sollte jeder seinen eigenen Kode definieren können. Statt den Interessierten eine Lösung ihrer Probleme zu geben, gab man ihnen die Möglichkeit, sich selbst eine Lösung zu bauen.

Die Flucht in die Metaebene gelingt nicht immer. Man verliert dabei leicht den Bezug auf die konkreten Probleme, die zu lösen sind. In diesem Fall aber scheint dies ein erfolgreiches Manöver gewesen zu sein.

In den Diskussionen um eine solche Metasprache hat sich der Begriff der Geräteunabhängigkeit allmählich etwas erweitert. Wenn man einmal gesehen hat, dass man „Pagethrow" schreiben kann, statt einer gerätespezifischen Anweisung, so sieht man bald auch ein, dass man ebenso „kursiv" schreiben könnte, um in eine geeignete kursive Schrift zu wechseln, statt genau die neue Schriftart mit Namen und Größe und Durchschuss anzugeben. Denn im folgenden Jahr will man vielleicht nicht mehr *Palatino* sondern *Garamond* haben; die Stellen, die kursiv (oder fett, oder gesperrt) zu setzen sind, bleiben aber die gleichen.

Bald ging man noch weiter. Statt anzugeben, dass gewisse Wortfolgen kursiv (bzw. fett oder gesperrt) zu setzen sind, kann man fragen, warum man diese Wortfolgen so setzt. Sie sind vielleicht *Termini technici* oder Schlüsselwörter; sie sind vielleicht Fremdwörter; sie sind vielleicht emphatisch, stark betont. Oder sie sind eben Kapitelüberschriften. Wenn man die typographisch

auszuzeichnenden Stellen nach Art identifiziert und diese Artangabe von den Gestaltungsanweisungen grundsätzlich trennt, dann vermindert man die Redundanz in der Datei (die Gestaltungsanweisungen für Kapitalüberschriften werden nur einmal angegeben), und es wird leichter, diese Anweisungen zu ändern.

Das Procedere ist das eines Makrosystems: Für jede zu unterscheidende Gestaltungsart definiert man ein Makro und man schreibt in der Datei nicht die Verarbeitungsanweisungen, sondern den Namen des Makros. Sehr früh hatten Formatierungsprogramme solche Makros angeboten. Die *GenCode* Arbeitsgruppe hat sich aber nicht damit begnügt, den Einsatz eines Makrosystems zu empfehlen. Sie haben scheinbar im Zusammenspiel von Makros, Anwendungsunabhängigkeit und Wiederverwendbarkeit eine tiefere Frage gesehen. In der Textverarbeitung funktionieren Makrosysteme am besten, wenn man sich nicht mehr fragt „Wie soll man diese Daten verarbeiten?", sondern „Was bedeuten diese Daten? Was sind diese Daten? Welche Art von Ding ist das? Was ist sein Genus?" Aus dieser letzten Frage stammt die Benennung *generisches Markup*: Markup, das das Genus einzelner Textstellen angibt.

Das generische Markup führt unwillkürlich und unaufhaltsam zur Ontologie.

Und zwar, sowohl im ingenieurtechnischen wie auch im philosophischen Sinne. Will man in einem Datenformat alle möglichen Applikationen, alle möglichen Softwares unterstützen (wie in der Rechtskanzlei sowohl die Dokumenterzeugung wie auch die Archivierung und die Volltextsuche), dann müssen zwei Strukturen X und Y in der Datei immer dann eindeutig unterscheidbar sein, wenn wenigstens eine der Applikationen diese Unterscheidung braucht, weil sie X und Y verschieden behandeln will. Die Strukturen (oder Objekte) X und Y dürfen das gleiche Genus haben, wenn jede zu unterstützende Applikation (im Grenzfall: jede denkbare Applikation) sie beide gleich behandeln soll. Das trifft dann zu wenn X und Y wesentlich gleicher Art sind, d. h. das gleiche Genus haben. Eine Folge dieser Beobachtung ist, dass jede Makrosammlung (in XML: jede Definition eines Dokumenttyps) explizit oder implizit eine Ontologie zur Schau stellt und dass ein Kode für die Dokumentverarbeitung eine Weltanschauung ausdrückt.

Dieser Schluss ist historisch und soziologisch gesehen sehr wichtig. Dass man selber als Benutzer, als Anwender, als Dateneigner den Kode definieren darf und muss, das bedeutet zuerst eine Art Selbstbestimmung wie man sie bei den Computern nicht immer so gewohnt war. Diese Selbstbestimmung bedeutet eine gewisse Freiheit, den Softwareentwicklern und Softwarelieferanten gegenüber. Die Selbstbestimmung bedeutet auch eine gewisse Verantwortung, denn man muss für sich selbst entscheiden, was die wesentlichen Züge der zu digitalisierenden Daten sind, welche Unterschiede als wesentlich zu gelten

haben, und welche als nichtwesentlich. Kant hat einmal geschrieben: „Aufklärung ist der Ausgang des Menschen aus seiner selbst verschulden Unmündigkeit." (Kant 1784) Und ein Dateneigner, der die Möglichkeit bekommt, selbst zu bestimmen, was in den Daten wesentlich ist, fühlt sich oft von einer Unmündigkeit befreit, denn es bestimmen nicht mehr unsichtbare Softwareentwickler was dem Benutzer in einem Text wichtig zu sein hat. Viele SGML- und XML-Anwender schätzen diese Selbstbestimmung sehr und wehren sich gegen jede Alternative, die wieder in die Unmündigkeit zu führen scheint.

3 Zentrale Konzepte von SGML and XML

Die Selbstbestimmung, so wichtig sie auch ist, ist nicht der einzige Grund, warum nach dreißig Jahren SGML und XML noch Anwendung finden. Einige technischen Eigenschaften von SGML und XML spielen auch eine Rolle.

Der Kern von SGML und XML besteht aus drei Ideen:

1. Die Syntax, die Definition der Zeichenfolgen, die für diese Formate wohl- bzw. nicht wohlgeformt sind.
2. Die Datenstrukturen, die mehr oder weniger implizit oder explizit in dieser Zeichenfolge immanent sind.
3. Die Möglichkeit, die Gültigkeit einer Zeichenfolge nachzuprüfen mittels einer Dokumentgrammatik.

Diese drei greifen ineinander ein und stärken sich gegenseitig. Sie bilden eine Art *Circulus*, aber keinen *Circulus vitiosus*, sondern ein *Circulus virtuosus*.

4 Syntax

Die Syntax von XML lässt sich sehr schnell beschreiben: Es wird alles explizit abgegrenzt.

Es soll so einfach wie möglich sein, mit einfachen Mitteln eine XML-Datei zu verarbeiten; bei der Entwicklung von XML hat man oft den *desperate Perl hacker* (den/die verzweifelte Perl-Programmierer/-in) als Zielpublikum besprochen: Es sollte ihm bzw. ihr möglich sein, die eigenen Dateien ohne großen Aufwand zu verarbeiten. Die explizite Abgrenzung aller XML-Strukturen im Text hilft dabei: Man kann sowohl den Anfang wie auch das Ende einer Struktur mit einfachen regulären Ausdrücken finden.

Jedes SGML- oder XML-Dokument besteht aus Elementen. Jedes Element wird durch Etiketten oder Tags von anderen Elementen abgegrenzt. Etiketten

werden gegen anderen Inhalt durch spitze Klammer abgegrenzt. Attributwerte, die in einem Starttag erscheinen, werden durch Anführungszeichen abgegrenzt.

Die Syntax von XML könnte ohne Informationsverlust wesentlich weniger langatmig und explizit sein: Anführungszeichen sind nur dann für Attributwerte dringend notwendig, wenn die Zeichenfolge ohne sie zweideutig wäre, oder schwierig zu parsen. Ende-Tags für Elemente könnte man oft ohne Informationsverlust auslassen: Da Elemente nisten, genügt der Ende-Tag eines äußeren Elements, das Ende aller noch nicht beendeten Elemente anzuzeigen, die innerhalb des äußeren Elements angefangen haben. SGML macht von solchen Möglichkeiten viel Gebrauch, XML zum großen Teil deswegen nicht, weil die Erfahrung gelehrt hat, dass solche Raffiniertheiten die einfache Textverarbeitung unverhältnismäßig erschweren, die Syntax komplizieren und die Wahrscheinlichkeit einer korrekten Implementierung vermindern (wie man am Beispiel von Mozilla und Internet Explorer leicht sieht: Jeder Implementierungsfehler der einen Software wurde schnell von der anderen in einer Abwärtsspirale nachgeahmt; die Syntax, die sie am Ende implementiert haben, hatte nur eine entfernte Ähnlichkeit mit der von der HTML-Norm vorgeschriebenen).

Als Beispiel der Syntax von XML diene hier ein Lied von Walther von der Vogelweide (Kuhn 1965, 52):

```
<Kanzone>
  <Strophe>
    <Aufgesang>
      <Stollen>
        <Z>“Unter den linden</Z>
        <Z>an der heide,</Z>
        <Z>dâ unser zweier bette was,</Z>
      </Stollen>
      <Stollen>
        <Z>dâ mugt ir vinden</Z>
        <Z>schone beide</Z>
        <Z>gebrochen bluomen unde gras.</Z>
      </Stollen>
    </Aufgesang>
    <Abgesang>
      <Z>vor dem walde in einem tal,</Z>
      <Z>tandaradei</Z>
      <Z>schône sanc diu nahtegal.</Z>
    </Abgesang>
  </Strophe>
```

```
  <Strophe>
    <Aufgesang>
      <Stollen>
        <Z>Ich kam gegangen</Z>
        <Z>zuo der ouwe:</Z>
        <Z>dô was mîn friedel komen ê.</Z>
      </Stollen>
      <Stollen>
        <Z>dâ wart ich empfangen</Z>
        <Z>hêre frouwe,</Z>
        <Z>daz ich bin sælic iemer mê.</Z>
      </Stollen>
    </Aufgesang>
    <Abgesang>
      <Z>kuster mich? wol tûsentstunt:</Z>
      <Z>tandaradei,</Z>
      <Z>seht wie rôt mir ist der munt.</Z>
    </Abgesang>
  </Strophe>
</Kanzone>
```

Hier kann man in den Tags sehen, was der Kodierer als das Wesentliche im Gedicht angesehen hat. Das Lied ist eine Kanzone; der Anfang wird mit dem Anfangstag <Kanzone> angezeigt, das Ende mit dem Ende-Tag </Kanzone>. Eine Kanzone besteht aus Strophen (hier durch <Strophe> … </Strophe> markiert). In der Kanzonenform besteht eine Strophe aus Teilen, die man später Aufgesang und Abgesang genannt hat. (Diese Termini kannten Walther und seine Zeitgenossen soweit wir wissen nicht, sie stammen von den Meistersängern.) Jeder Aufgesang besteht aus zwei Stollen, die aus Zeilen bestehen, wie auch der Abgesang. Alle diese Strukturen werden explizit mit XML-Elementen modelliert und können leicht von einem Programm bearbeitet werden. Es wäre mit einer Gedichtsammlung in dieser Form z. B. leicht, nachzuprüfen, ob der Wortschatz oder das Reiminventar des Aufgesangs von dem Wortschatz und dem Reiminventar des Abgesangs bei Walther oder allgemein unterscheidet.

Diese Begriffe – Strophe, Aufgesang, Stollen – sind für das Verständnis der Liedform wesentlich, aber man kann nicht erwarten, sie bei einem kommerziellen Wortprozessor wie Microsoft Word vordefiniert zu finden. Ähnliches gilt in jedem Bereich, wo man sich intensiv mit Texten befasst: Wichtige fachspezifische Begriffe lassen sich erst mit deskriptivem Markup im Text markieren.

5 Datenmodelle

Die zweite Kernidee von SGML und XML sind die Datenmodelle. Streng genommen definieren die SGML-Norm und die XML-Spezifikation nur die Menge der konformen Zeichenfolgen; sie bieten keine explizite Beschreibung eines Datenmodells. Das steht in einem starken Kontrast zu der Entwicklung der Datenbanktheorie, die sowohl ein Datenmodell explizit definiert wie auch die Schnittstelle, durch die ein Programmierer modellgemäß auf die Daten zugreifen kann. Man hat oft genug das Fehlen eines Modells bei XML als Mangel betrachtet.

Ein explizites Modell ist wichtig, wenn man eine Schnittstelle definiert. XML und SGML wollen aber keine Schnittstellen definieren: Sie definieren ein Datenformat. Implizit darf man verschiedene Modelle definieren und auf dieses Format anwenden.

Als Zeichenfolge: SGML- und XML-Dokumente werden als Zeichenfolgen definiert: man kann also jedes Dokument als Zeichenfolge auffassen. Jede Software, die Daten als Zeichenfolge modelliert, kann erfolgreich auf SGML- und XML-Dateien verwendet werden. Das erlaubt uns, SGML- und XML-Dateien mit dem systemüblichen Kopierprogramm zu kopieren; man bedarf nicht (wie bei vielen relationellen Datenbanken) eines modellspezifischen Kopierkommandos.

Als reguläre Abwechslung von Markup und Inhalt: Die Zeichenfolge einer SGML- oder XML-Datei wird durch eine Grammatik definiert, die zwischen Zeichendaten (Dokumentinhalt) und Markup (Tags u. a. m.) unterscheidet. Man kann demnach die Zeichenfolge als eine Abwechslung zwischen Zeichendaten und Markup betrachten und verarbeiten. Tags lassen sich leicht mit regulären Ausdrücken erkennen.

Als Baum: Aus der Verschachtelung der Elemente ergibt sich, dass SGML- und XML-Dokumente leicht mit einer Baumstruktur modelliert werden können. Einen Teil der Baumstruktur des Waltherlieds sieht man in Abbildung 13.1.

Als gerichteter Graph: Mithilfe der Document Type Definition (s. weiter unten) kann man bestimmte Attributwerte als Zeiger definieren, und so von einem XML-Element auf ein anderes verweisen. Wenn man diese Verweise in Betracht zieht, dann hat man mit keinem Baum mehr zu tun, sondern mit einem gerichteten Graph. Wenn man die oben gezeigte Baumstruktur um Verweise erweitert, die Reimpaare verbinden, erhält man den Graph in Abbildung 13.2.

Als anwendungsspezifische Datenstruktur: Ein Programm, das eine XML-Datei verarbeitet, kann und darf sich eine eigene Datenstruktur nach Belieben aus den Informationen der Datei bauen. Allgemeiner als ein gerichteter Graph wird diese Struktur nicht sein, aber man ist nicht daran gebunden, nur die im XML-Markup explizit angegebene Strukturen in Betracht zu ziehen. Für viele Anwendungen genügen aber die oben aufgeführten Datenstrukturen und man kann sich den Bau einer eigenen Datenstruktur ersparen.

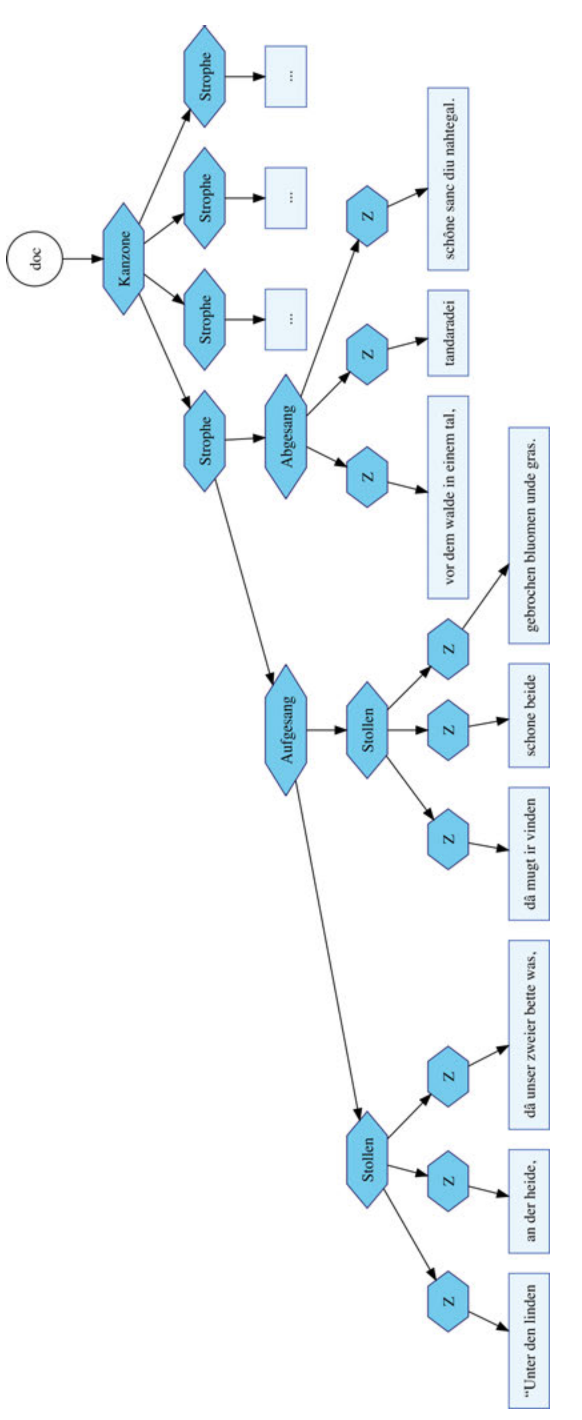


Abb. 13.1: Baumstruktur des Waltherliedes

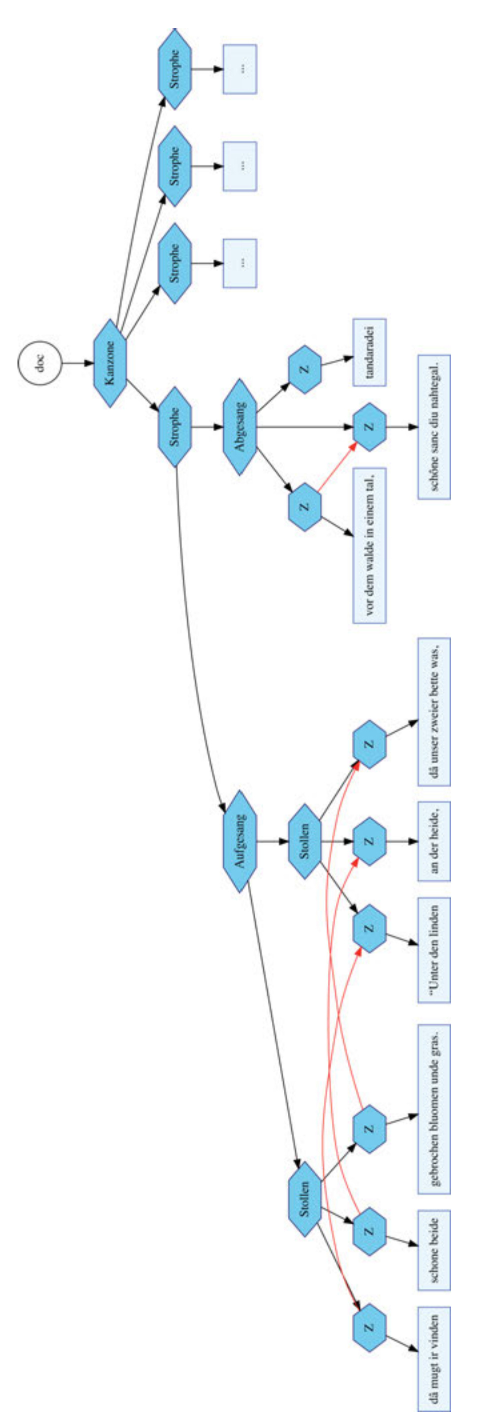


Abb. 13.2: Waltherlied als gerichteter Graph.

6 Dokumentgrammatik

Die dritte Kernidee von SGML und XML ist die Dokumentgrammatik.

Was man als Baum betrachten kann, kann man auch als Syntaxbaum betrachten. Und ein Syntaxbaum wird von einer Grammatik erzeugt.

SGML und XML erlauben dem Anwender nicht nur, eigene Elementtypen zu gebrauchen, sondern auch die wechselseitigen Beziehungen der Elementtypen. Als Beispiel diene eine Dokumentgrammatik für das oben gezeigte Lied Walthers.

```
<!ELEMENT Kanzone (Strophe+)>
<!ELEMENT Strophe (Aufgesang, Abgesang)>
<!ELEMENT Aufgesang (Stollen, Stollen)>
<!ELEMENT Stollen (Z+)>
<!ELEMENT Abgesang (Z+)>
<!ELEMENT Z (#PCDATA)>
```

Diese Elementvereinbarungen schreiben vor, welche Elementtypen wo erscheinen dürfen. Ein Element Kanzone besteht aus einer Strophe oder aus mehreren Strophen. Das Element Strophe besteht aus Aufgesang und Abgesang, beides obligatorisch. Ein Aufgesang besteht aus zwei Stollen. Ein Stollen besteht aus Zeilen (Z), was auch auf den Abgesang zutrifft. Und eine Zeile (ein Element vom Typ Z) besteht aus einer Zeichenfolge (#PCDATA).

Man sieht, dass es möglich ist, mittels dieser Dokumentgrammatik einen großen Teil der Definition der Kanzonenform festzuhalten, wie man auch einen großen Teil der Syntax einer Programmiersprache mittels einer kontextfreien Grammatik festhalten kann. In den meisten Programmiersprachen gibt es aber Vorschriften, die man in einer kontextfreien Grammatik nicht beschreiben kann; sie überschreiten die Aussagekraft des Formalismus. So auch hier: Die Dokumentgrammatiken von SGML sind im Grunde leicht eingeschränkte kontextfreie Grammatiken. Es ist zum Beispiel nicht möglich, in einer Dokumenttypdefinition vorzuschreiben, dass die zwei Stollen die gleiche Länge haben müssen; die Länge eines Stollens wird nicht vorgeschrieben, aber die zwei Stollen müssen die gleiche Länge haben. Und der Abgesang muss länger als ein Stollen sein, aber kürzer als der Aufgesang. Diese Einschränkungen sind mit den Dokumenttypdefinitionen (DTDs) von SGML und XML nicht auszudrücken. (Es gibt inzwischen aber andere Schemasprachen, die es erlauben, diese Bedingungen zu formulieren; dass auch sie nicht alle denkbaren Bedingungen berechnen können, folgt aus den Arbeiten Gödels und Turings.)

Wie jede Grammatik kann eine Dokumentgrammatik eng oder breit geschrieben werden, kann deskriptiv eine bereits existierende Menge von Dokumenten beschreiben oder präskriptiv vorschreiben, wie eine gewisse Art von Dokumenten zu strukturieren sei. In vielen Wörterbuchverlagen möchte man eine rein präskriptive Grammatik für die Wörterbucheinträge haben: Die Einträge haben eine komplexe Struktur, und es geschieht leicht, dass man aus Versehen den eigenen Strukturregeln nicht folgt. Da eine Dokumentgrammatik eine rein automatische Überprüfung der Struktur ermöglicht, kann man mit einer präskriptiven Grammatik eine weit höhere Konsistenz der Einträge erreichen. Aber diese präskriptive, vorschreibende Grammatik kann man praktisch auf die bereits erschienenen Wörterbücher nie anwenden, denn es gibt in jedem veröffentlichten Wörterbuch Strukturvarianten, die der Verlag als Fehler betrachtet. Die Möglichkeit, in Zukunft solche Fehler zu verhindern, macht SGML und XML für viele praktische Anwendungen allen Alternativen konkurrenzlos überlegen.

7 Schlusswort

Diese drei Ideen, und ihre Verwirklichung in SGML und XML, wurden von vielen Anwendern als Revolution empfunden. Dass Dokumente eine Struktur haben, das weiß jeder, der mit Dokumenten täglich arbeitet. Dass diese Dokumentstruktur ohne großen Aufwand der Anwendungssoftware zugänglich gemacht werden soll, leuchtet sofort ein. Und dass der Benutzer selbst bestimmen kann, wie man die Elemente benennt und wie die Strukturen aufzubauen sind, gibt dem Benutzer von SGML und XML eine Art Freiheit, die man nur mit Adam im Garten Eden vergleichen kann, als Adam und Eva alle Tiere der Welt nach ihrer Art benannt haben. Man definiert sich eine Welt damit.

Die Einführung der kontextfreien Grammatik in *Backus/Naur Form* (BNF) bei der Definition von *Algol 60* gilt als Wendepunkt in der Geschichte der Programmiersprachen. Sie hat ein bis dahin nicht erreichbares Niveau an Formalität und Eindeutigkeit bei der Syntaxdefinition der Programmiersprachen möglich gemacht. Die Welt der Programmiersprachen ist eine andere geworden. Die Welt der Dokumentverarbeitung ist im Vergleich vielleicht nur eine kleine Welt, aber diese Welt ist auch infolge der Kernideen von SGML und XML eine andere und bessere geworden.

Literatur

International Organization for Standardization (ISO) (1986): *ISO 8879-1986 (E), Information processing – Text and office systems – Standard Generalized Markup Language (SGML)*. Geneva: International Organization for Standardization.

International Organization for Standardization (ISO) (2016): *ISO 9075–14:2016, Information technology – Database languages – SQL – Part 14: XML-related specifications (SQL/XML)*. Geneva: International Organization for Standardization.

Kant, Immanuel (1784): Beantwortung der Frage: Was ist Aufklärung? Berlinische Monatsschrift, Dezember 1784, 481–494. Häufig nachgedruckt, u. a. in: Immanuel Kant: Wast ist Aufklärung? Ausgewählte kleine Schriften. Hsg. Von Horst D. Brandt. Hamburg: Felix Meiner Verlag, 1999 [= Philosophische Bibliothek Band 512], 20–26.

Kuhn, Hugo (1965): Die Gedichte Walthers von der Vogelweide. Hrsg. von Karl Lachmann. Dreizehnte, aufgrund der zehnten, von Carl von Kraus bearbeiteten Ausgabe, neu herausgegeben von Hugo Kuhn. Berlin: de Gruyter.

Library of Congress (LC) (2016): *MARCXML: MARC 21 XML Schema*. Washington, DC: Library of Congress 6. http://www.loc.gov/standards/marcxml/ (letzter Zugriff: 6. 11. 2017).

Library of Congress (LC) (2017a): *METS: Metadata Encoding & Transmission Standard.* Washington, DC: Library of Congress. http://www.loc.gov/standards/mods/ (letzter Zugriff: 6. 11. 2017).

Library of Congress (LC) (2017b): *MODS: Metadata Object Description Schema.* Zuletzt v.3.6. Washington, DC: Library of Congress. http://www.loc.gov/standards/mods/ (letzter Zugriff: 6. 11. 2017).

Lüngen, Harald & C. M. Sperberg-McQueen (2012): A TEI P5 Document Grammar for the IDS Text Model. In Piotr Bański, Eleonora Modignani Picozzi & Andreas Witt (Hrsg.), *Journal of the Text Encoding Initiative* 3, *Special Issue on TEI and Linguistics*. http://jtei.revues.org/508 (letzter Zugriff: 6. 11. 2017).

TEI Consortium (TEI) (2017): *TEI P5: Guidelines for Electronic Text Encoding and Interchange, version 3.2.0*. o. O.: TEI Consortium. http://www.tei-c.org/Guidelines/P5/ (letzter Zugriff: 6. 11. 2017).

World Wide Web Consortium (W3C) (1998): *Extensible Markup Language (XML) 1.0*. First edition: Tim Bray, Jean Paoli & C. M. Sperberg-McQueen (Hrsg.). Fifth Edition (2008), W3C Recommendation 26 November 2008: Tim Bray et al. (Hrsg.). Cambridge, Sophia-Antipolis, Tokyo: W3C. http://www.w3.org/TR/REC-xml (letzter Zugriff: 6. 11. 2017).

World Wide Web Consortium (W3C) (2010): *XQuery 1.0: An XML query language*. First edition 14 December 2010: Scott Boag et al. (Hrsg.). Zuletzt: Jonathan Robie, Michael Dyck & Josh Spiege (Hrsg.), *XQuery 3.1: An XML query language, W3C Recommendation 21 March 2017.* Cambridge et al.: W3C. http://www.w3.org/TR/Rec-xquery-20101214/; http://www.w3.org/TR/xquery-31 (letzter Zugriff: 6. 11. 2017).

World Wide Web Consortium (W3C) (2011): *Scalable Vector Graphics (SVG) 1.1* (Second Edition), W3C Recommendation 16 August 2011: Erik Dahlström et al. (Hrsg.). Cambridge, Sophia-Antipolis, Tokyo: W3C.

Michael Beißwenger

14 Internetbasierte Kommunikation und Korpuslinguistik: Repräsentation basaler Interaktionsformate in TEI

Abstract: Der Beitrag beschreibt ein Basisschema für die Repräsentation von Korpora internetbasierter Kommunikation auf der Grundlage der *Guidelines for Electronic Text Encoding and Interchange*“ der *Text Encoding Initiative* (TEI). Ausgehend von einem Überblick über die gegenwärtige Korpuslandschaft wird gezeigt, dass sich in der Kommunikation im Netz trotz des beständigen technologischen Wandels stabile Interaktionsformate etabliert haben. Eine standardisierte Repräsentation solcher Formate bildet eine wichtige Voraussetzung für die Herstellung einer sprachen- wie domänenübergreifenden Interoperabilität von Korpora und leistet einen Beitrag zum Aufbau der Sprachressourcen-Infrastruktur der Zukunft.

Keywords: Digital Humanities, Interaktion, Internetbasierte Kommunikation, Sprachkorpora, TEI

> *„Das Internet? Gibt’s diesen Blödsinn immer noch?“*
> (Homer Simpson)

1 Einleitung

Seit 2013 befasst sich eine Arbeitsgruppe (*special interest group*) im Rahmen der *Text Encoding Initiative* (TEI) unter dem Titel *Computer-mediated communication* (kurz: TEI-CMC-SIG) mit der Entwicklung eines XML-Schemas für die Repräsentation und Strukturannotation von Sprachdaten aus Formen internetbasierter Kommunikation. Das von der Gruppe entwickelte Schema soll eine einheitliche und softwareunabhängige, texttechnologische Modellierung von Sprachdaten internetbasierter Kommunikation in linguistischen Korpora ermöglichen und sich zugleich als Austauschformat für Korpusdaten eignen. Das Schema ist als

Michael Beißwenger, Universität Duisburg-Essen, Institut für Germanistik, Berliner Platz 6–8, D-45127 Essen, E-Mail: michael.beisswenger@uni-due.de

https://doi.org/10.1515/9783110538663-015

ein *Basis*schema konzipiert, dessen Grundstruktur sich aus Strukturinformationen, die in den Korpus-Ausgangsdaten (die beispielsweise im HTML-Format vorliegen) bereits enthalten sind, durch einfache Transformationen möglichst automatisch erzeugen lassen soll. Die Motivation für die Erarbeitung eines solchen Basisschemas zielt auf die Bereitstellung einer Lösung für die Aufgabe der Strukturannotation, die für viele Korpusprojekte (zu unterschiedlichen Sprachen und für Daten aus unterschiedlichen IBK-Formen) praktikabel ist und die sich projektspezifisch erweitern lässt. Die Entscheidung, ein solches Schema auf der Grundlage des Encoding-Standards der TEI zu entwickeln, ist motiviert durch dessen Flexibilität: TEI-Schemas lassen sich über den Mechanismus der *customization* erweitern und an individuelle Bedürfnisse anpassen.[1] Daneben ist die Entscheidung für TEI aber auch dadurch motiviert, dass die von der TEI bereitgestellten Formate im Bereich der Digital Humanities als ein *De-facto*-Standard etabliert sind: Viele Sprachressourcen und Korpora – zum Beispiel das Deutsche Referenzkorpus DeReKo am Institut für Deutsche Sprache (IDS) und die DWDS-Textkorpora an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) – nutzen diese Formate bereits (Lüngen & Sperberg-McQueen 2012, CLARIN-D User Guide 2012: Kap. 6). Wer sich dafür entscheidet, seine Sprachressourcen in TEI zu repräsentieren, der sichert seinen Ressourcen damit eine grundsätzliche *Interoperabilität* mit anderen Sprachressourcen: Repräsentationsformate, die von vielen verwendet werden und die darüber hinaus – wie im Falle der TEI-Formate – softwareunabhängig sind, werden mit sehr hoher Wahrscheinlichkeit auch in 20 Jahren noch gepflegt und von gängigen korpustechnologischen Werkzeugen verarbeitet werden können (*Nachhaltigkeit*). Interoperabilität vereinfacht darüber hinaus aber auch die Vernetzung von Ressourcen unterschiedlicher Urheber (*Kombinierbarkeit*). Ressourcen, die auf zumindest basaler Ebene in einem einheitlichen XML-Format repräsentiert sind, lassen sich mit weniger technischem und konzeptionellem Aufwand zusammenführen und mit den gleichen Werkzeugen vergleichend auswerten als Ressourcen, die völlig unterschiedliche Formate verwenden. Nicht zuletzt verringert der Rückgriff auf Standards den Aufwand beim Aufbau neuer Ressourcen: Standards dienen dazu, „dass nicht jeder das Rad neu erfinden muss“ (Lobin 2010: 107); sie ermöglichen, dass für die zu lösende Aufgabe (im hier besprochenen Fall die Aufgabe der Repräsentation und Strukturannotation von Korpora internetbasierter Kommunikation) auf Lösungen zurückgegriffen kann, die sich in anderen Projekten ähnlicher

1 Siehe hierzu das Kapitel *Customization* in den TEI-Guidelines: http://www.tei-c.org/Guidelines/Customization/index.xml (letzter Zugriff: 8. 11. 2017).

Art bereits bewährt haben (*Ökonomie* und *Praktikabilität*). Die Idee der Standardisierung zielt darauf, das Gemeinsame zu erfassen und einheitlich zu beschreiben, das für einen Gegenstand bzw. eine Domäne übergreifend zu den spezifischen Anforderungen und Forschungsinteressen in einzelnen Projektzusammenhängen festgestellt werden kann. Standardisierung und das, was ein Standard abbilden kann (und sollte), hat damit notwendigerweise Grenzen (Perkuhn, Keibel & Kupietz 2012: 68). Gerade deshalb zielt die Arbeit der TEI-CMC-SIG auf die Entwicklung lediglich eines Basisschemas, das grundlegende Struktureigenschaften erfasst, deren Beschreibung in vielen Projekten von Interesse ist. Auf dieses Basisschema sollen weitere, projektspezifische Annotationen aufsetzen können.

In ihrer gegenwärtigen Fassung P5 enthalten die TEI-Guidelines noch keine Modelle für die Strukturbeschreibung von Formen internetbasierter Kommunikation. Die Entwicklung von Lösungen erfolgt daher gegenwärtig auf der Ebene von *customizations*. *Customization* bezeichnet eine Strategie bei der Entwicklung von TEI-Schemas, die es erlaubt, TEI in Domänen anzuwenden, die der Standard in seiner aktuellen Version noch nicht erfasst:

> Because the TEI Guidelines must cover such a broad domain and user community, it is essential that they be customizable: both to permit the creation of manageable subsets that serve particular purposes, and also to permit usage in areas that the TEI has not yet envisioned. Customization is a central aspect of TEI usage and the Guidelines are designed with customization in mind. (TEI-P5: Customization; Hervorhebung MB)

Bislang liegen drei *customizations* für die Strukturannotation von IBK-Korpora vor, die im Zusammenhang mit der Arbeit der TEI-CMC-SIG entwickelt und die bereits in verschiedenen Korpusprojekten zum Deutschen und zum Französischen eingesetzt werden. Nächste Schritte der Arbeit der TEI-CMC-SIG werden darin bestehen, ausgehend von den entwickelten *customizations* eine Eingabe in den Standardisierungsprozess der TEI zu formulieren, um Modelle für die Strukturannotation von IBK-Korpora als echte Elemente einer künftigen Version des TEI-Standards zu verankern.

Der vorliegende Beitrag beschreibt die Zielsetzung und den Ausgangspunkt sowie den aktuellen Stand der Schemaentwicklung der TEI-CMC-SIG. Dazu wird zunächst ein Überblick über die Abdeckung internetbasierter Kommunikation in der gegenwärtigen Korpuslandschaft und über Desiderate im Zusammenhang mit dem Aufbau von Korpora internetbasierter Kommunikation gegeben (Abschnitt 2). Anschließend wird die Frage diskutiert, inwieweit die internetbasierte Kommunikation, die durch beständigen technologischen Wandel geprägt ist, überhaupt sinnvoll in einem Repräsentationsformat abgebildet werden kann, das beansprucht, stabile *Interaktions*formate abzubilden

(Abschnitt 3). Abschnitt 4 gibt einen Überblick über den in 2017 erreichten Stand der Schemaentwicklung. Der Beitrag schließt mit einem Ausblick auf künftige Arbeiten und auf die Korpuslandschaft von morgen (Abschnitt 5).

2 Internetbasierte Kommunikation und Korpuslinguistik: Lagebericht, Forschungsbaustellen und Projekte

Die wissenschaftliche Beschäftigung mit der Kommunikation im Internet kann auf eine gut 25-jährige Geschichte zurückblicken. Einen Fokus der linguistischen Forschung zum Gegenstand bildet dabei von Anfang an die Spezifik dialogisch-interaktionaler Sprachverwendung unter den Bedingungen digitaler Vermittlung. Den so charakterisierten Forschungsgegenstand bezeichne ich als *Internetbasierte Kommunikation* (kurz: IBK).[2] Internetbasierte Kommunikation vollzieht sich in einer Vielzahl unterschiedlicher Kommunikations*formen*, die auf der Verwendung von unterschiedlichen Kommunikations*technologien* beruhen, die je spezifische mediale und modale Ressourcen sowie Bedingungen für die Produktion und Rezeption von kommunikativen Äußerungen vorgeben.[3] Dazu zählen einerseits Technologien, die für die Nutzung anhand von Browsersoftware oder von Clientanwendungen auf dem PC konzipiert sind, und andererseits Technologien, die über sogenannte Apps auf Smartphones und Tabletcomputern als Endgeräten bereitgestellt werden. Eine Kommunikations-

---

2 Zum Gegenstandsbereich gibt es in der Forschungsliteratur unterschiedliche Vorschläge zur terminologischen Konzeptualisierung, die letztlich einen hohen Grad an extensionaler Übereinstimmung aufweisen. Am ältesten und verbreitetsten ist die Etikettierung als *computer-mediated communication* (CMC, z. B. Herring 1996), ins Deutsche lehnübersetzt als *Computervermittelte Kommunikation*. Der hier präferierte Terminus *Internetbasierte Kommunikation* (IBK), der um die Jahrtausendwende als zeitgemäßere Alternative zu CMC geprägt und u. a. im DFG-Netzwerk *Empirikom* (Beißwenger 2017) verwendet wurde, grenzt die Kommunikation auf Basis von TCP/IP von anderen Formen computervermittelter Kommunikation ab (auch Briefe und Telefongespräche werden heutzutage unter Beteiligung von Computern vermittelt). Jucker & Dürscheid (2012) schlagen die Bezeichnung *Keyboard-to-screen-Kommunikation* vor, die die Spezifik der Ein-/Ausgabedimension fokussiert. *Cum grano salis* werden unter allen diesen Etiketten Erscheinungsformen interaktionaler Sprache unter den Bedingungen digitaler Vermittlung auf Basis von Computernetzwerken untersucht. Für eine eingehendere Diskussion der verschiedenen Konzeptualisierungen sei auf Storrer (2018) verwiesen.

3 Zur Unterscheidung von Kommunikationstechnologien und Kommunikationsformen vgl. Beißwenger (2007: 107–112).

technologie kann dabei exklusiv für den Betrieb einer einzigen Anwendung konzipiert sein, deren Anbieter zugleich der Entwickler[4] der Technologie ist; Beispiele hierfür sind die Technologien, auf denen Social-Network-Plattformen wie Facebook, Google+, Twitter oder Instagram, die Video-Plattform YouTube, Lernplattformen wie moodle und stud.IP oder Kommunikationsdienste wie WhatsApp, Threema, Snapchat oder Skype beruhen. Andere Kommunikationstechnologien existieren in Form von Software, die für die Installation und den Betrieb durch potenziell beliebig viele Anbieter bereitgestellt wird. Dazu zähle ich u. a. E-Mail-, Foren-, Chat-, Weblog- und Wiki-Software, die kostenfrei oder gegen Entgelt für eigene Instanziierungen entsprechender Anwendungen genutzt werden kann. Kommunikation zwischen Nutzern zu ermöglichen, kann die primäre Funktion einer Kommunikationstechnologie sein (Chats, Foren, WhatsApp, Skype) oder auch nur eine Funktion neben weiteren darstellen (wie z. B. im Falle von sozialen Netzwerken, Lernplattformen, Wikis und YouTube). Innerhalb von Kommunikations*formen* sind verschiedene kommunikative *Gattungen* realisierbar – vgl. hierzu analog die Unterscheidung der Konzepte „Kommunikationsform" und „Textsorte" bei Brinker, Cölfen & Pappert (2014: 140–142) –, speziell in Bezug auf die internetbasierte Kommunikation entstehen dabei auch neue Gattungen, die zwar Vorläufer, aber keine direkten Entsprechungen im Bereich der gesprochenen Sprache bzw. der nicht-internetbasierten Kommunikation haben. Kandidaten für solche neuen Gattungen sind in Chats z. B. die „interaktiven Lesespiele" (Beißwenger & Storrer 2012), in Online-Computerspielen die multimodalen „Raids" (Stertkamp 2016), in sozialen Netzwerken die sogenannten „Cybermobbing-Diskurse" (Marx 2015).[5]

Seit ihren Anfängen um Anfang der 1990er Jahre ist die IBK-Forschung stark empirisch ausgerichtet. Die starke empirische Ausrichtung steht aber – bis heute – in einem deutlichen Missverhältnis zur Verfügbarkeit frei zugänglicher Korpora, die für die linguistische Recherche und Analyse aufbereitet sind und die – wie das im Bereich der Textkorpora (z. B. Lüngen 2017; Geyken et al. 2017) bereits möglich ist – als Referenzressourcen für unterschiedlichste Arten von Forschungsfragen genutzt werden können. In aller Regel muss, wer ein sprachliches oder interaktionales Phänomen in der internetbasierten Kommunikation empirisch untersuchen möchte, ein geeignetes Datenset für sein Vorhaben selbst aufbauen. Das kostet Zeit, konzeptionelle Arbeit und häufig auch

4 Aus Gründen der besseren Lesbarkeit wird in diesem Beitrag für alle Personenbezeichnungen das generische Maskulinum gewählt. Die weibliche Form wird dabei stets mitgedacht. Sämtliche Personenbezeichnungen schließen somit beiderlei Geschlecht ein.

5 Zur Abgrenzung von Kommunikationsformen und Gattungen in Bezug auf die internetbasierte Kommunikation vgl. auch Dürscheid (2005).

Geld – Ressourcen, die sich gewinnbringender einsetzen ließen, wenn frei nutzbare Referenzkorpora zum Gegenstand existierten, die verschiedene Kommunikationsformen und -kontexte abdecken und somit zwar sicherlich nicht für alle, aber zumindest für eine große Zahl von Forschungsfragen als Datengrundlage herangezogen werden können.

Die unzureichende Abdeckung der internetbasierten Kommunikation in der aktuellen Korpuslandschaft ist nicht nur für die IBK-Forschung im engeren Sinne ein Problem, sondern generell für jedwede Forschung und praktische linguistische Tätigkeit, die auf die Beschreibung der deutschen Gegenwartssprache zielt: Angesichts der Bedeutung internetbasierter Kommunikation in vielen Bereichen des beruflichen und privaten Alltags und mit Blick auf die schiere Menge an geschriebener und gesprochener Sprache, die tagtäglich unter Nutzung internetbasierter Kommunikationstechnologien produziert wird, ist der Gegenstand „Deutsche Gegenwartssprache" in gegenwartssprachlich ausgerichteten Korpora und Korpussammlungen nur unzureichend erfasst, solange die Sprachverwendung in der internetbasierten Kommunikation darin nur unzureichend abgedeckt ist. Das schließt auch die lexikographische und grammatikographische Beschreibung der deutschen Gegenwartssprache und ihrer Varianten ein.

Um die Abdeckung der internetbasierten Kommunikation in der Korpuslandschaft zum Deutschen und zu anderen Sprachen zu verbessern, müssen die Korpuslinguistik und ihre Nachbardisziplinen (insbesondere die Sprach- und Texttechnologie) sich verschiedener Desiderate und Problembereiche annehmen, die sich im Zusammenhang mit der Erhebung, der Dokumentation, der Aufbereitung und der Wiederbereitstellung von IBK-Daten in Korpora stellen. Diese Desiderate sind schon länger bekannt (vgl. die Problemaufrisse in Beißwenger & Storrer 2008; Storrer 2014: 187–191; Bolander & Locher 2014; Lemnitzer & Zinsmeister 2015: 151–152). Entwicklungs- und Klärungsbedarf besteht insbesondere in Bezug auf die folgenden Fragenkomplexe:

- Fragen der Bewertung des rechtlichen Status für die Erhebung und Bereitstellung von IBK-Daten (exemplarisch iRights.Law Rechtsanwälte 2016; Beißwenger et al. 2017b), damit verbunden die Entwicklung von Konzepten und Verfahren für die Anonymisierung und Pseudonymisierung von IBK-Daten (hierzu u. a. DiDi 2015; Beißwenger et al. 2017b).
- Fragen der Erhebung von IBK-Daten aus Domänen privater Kommunikation (Facebook, WhatsApp, SMS ...) (Dürscheid & Stark 2011; Frey, Stemle & Glaznieks 2014; Verheijen & Stoop 2016).
- Entwicklung von Formaten für die Repräsentation von IBK-Daten – als Voraussetzung für den Austausch und die Vernetzung von Ressourcen (Beißwenger et al. 2012; Chanier et al. 2014).

- Entwicklung von Formaten für die Erfassung und Repräsentation von Metadaten; hier sind u. a. auch solche Metadaten zu berücksichtigen, die die Kommunikationsumgebungen beschreiben, aus denen die Korpusdaten erhoben wurden. Mit Blick auf die Veränderlichkeit von digitalen Technologien und Kommunikationsumgebungen sind für die Nachnutzbarkeit der Korpusdaten und für das Verständnis der mit diesen Daten dokumentierten Interaktionen 10, 20 oder 30, unter Umständen sogar bereits 3 Jahre nach ihrer Erhebung solche beschreibenden Daten von großer Wichtigkeit.
- Fragen der Anpassung sprachtechnologischer Verfahren für die Segmentierung und Klassifikation von Sprachdaten (Tokenisierung, morphosyntaktische Annotation, Lemmatisierung, Parsing) an die sprachlichen Besonderheiten internetbasierter Kommunikation (u. a. Giesbrecht & Evert 2009; Horbach et al. 2014; Horsmann & Zesch 2015; WAC-X & EmpiriST 2016).
- Anpassung von Korpusverwaltungs- und -abfragesystemen an die Anforderungen und Nutzerinteressen in Bezug auf IBK-Daten.

Die Gelegenheit für die Erarbeitung von projektübergreifend nützlichen Lösungen für die skizzierten Desiderate – und damit für die Entwicklung von Standards *bottom-up* – ist aktuell günstiger denn je: In den vergangenen Jahren wurden für verschiedene Sprachen und IBK-Formen Korpusprojekte auf den Weg gebracht, deren Ergebnisse nach Abschluss der Projektlaufzeit der Scientific Community als Ressourcen zur Verfügung gestellt werden sollen oder bereits zur Verfügung stehen. Die Erfahrungen und Lösungsansätze, die in diesen Projekten entwickelt werden, stellen wertvolle Ressourcen für die Entwicklung von projektübergreifenden Lösungen dar.[6]

Beispiele für aktuelle (abgeschlossene und laufende) Korpusprojekte sind:

- *CoMeRe*: eine Sammlung von vierzehn IBK-Korpora zum Französischen, die neun verschiedene IBK-Genres abdeckt (SMS, Wikipedia-Diskussionen, Tweets, Weblogs, E-Mails, Diskussionsforen, Chat, multimodale Formen) und die neben rein schriftlichen auch multimodale, synchrone und asynchrone Formen sowie Kommunikation aus dem öffentlichen und aus dem privaten Bereich erfasst. Sämtliche Korpora sind in TEI repräsentiert (Chanier et al. 2014) und werden unter einer CC-BY-Lizenz als OpenData über ORTOLANG[7] als downloadbare Ressourcen zur Verfügung gestellt.

---

6 Beispiele für Lösungsansätze aus verschiedenen Projekten beschreiben Beißwenger et al. (2017a).

7 http://hdl.handle.net/11403/comere (letzter Zugriff: 8. 11. 2017).

- *CorCenCC-CMC*: eine seit 2016 im Aufbau befindliche *e-language*-Komponente im Projekt *Corpws Cenedlaethol Cymraeg Cyfoes* (*National Corpus of Contemporary Welsh*, CorCenCC) mit Sitz an der Cardiff University (UK).[8]
- *DeReKo-News*: das seit 2013 aufgebaute deutsche Newsgroup-Korpus in DeReKo[9] im Umfang von 98 Millionen Tokens mit Daten aus den Jahren 2013–2016 (Schröck & Lüngen 2015).
- *DeReKo-Wikipedia*: die Wikipedia-Korpora in DeReKo, die Artikel- und Diskussionsseiten im Umfang von 581 Millionen Tokens enthalten und vom IDS sowohl als downloadbare Ressourcen als auch zur Abfrage über COSMAS II[10] angeboten werden (Margaretha & Lüngen 2014).
- *DiDi*: Korpus zur Sprachverwendung in Facebook mit Daten deutscher und italienischer Sprache sowie in Südtiroler Mundart, aufgebaut im Rahmen des DiDi-Projekts (*Digital Natives – Digital Immigrants*) an der EURAC in Bozen (IT) und online abfragbar via ANNIS (Frey, Glaznieks & Stemle 2016).[11]
- *Dortmunder Chat-Korpus*: Korpus mit einer Million laufenden Wortformen zur deutschsprachigen Chat-Kommunikation in unterschiedlichen Handlungsbereichen (Freizeit, Bildung, Beratung, Medien), die um XML-Annotationen zu ausgewählten Sprachmerkmalen angereichert wurden (Beißwenger 2013). Das Korpus wurde seit 2005 in einer „Releaseversion" zum freien Download angeboten.[12] Eine um *Part-of-speech*-Annotationen sowie eine TEI-Repräsentation erweiterte, vollständig anonymisierte Version (*Chat-Korpus 2.0*) wurde 2015–2017 in die CLARIN-D-Korpusinfrastrukturen integriert und steht seit Herbst 2017 als Teil des *DeReKo* und der Korpussammlung der Berlin-Brandenburgischen Akademie der Wissenschaften für die Online-Abfrage über die Korpus-Recherceschnittstellen am Institut für Deutsche Sprache (IDS) Mannheim und über das Portal www.dwds.de zur Verfügung (Lüngen et al. 2016; Beißwenger et al. 2017b).[13]

8 http://sites.cardiff.ac.uk/corcencc/ (letzter Zugriff: 8.11. 2017).
9 http://www.ids-mannheim.de/dereko (letzter Zugriff: 8.11. 2017).
10 https://cosmas2.ids-mannheim.de/ (letzter Zugriff: 8.11. 2017).
11 http://www.eurac.edu/didi (letzter Zugriff: 8.11. 2017).
12 http://www.chatkorpus.tu-dortmund.de/ (letzter Zugriff: 8.11. 2017).
13 Am IDS ist das Chat-Korpus über COSMAS II recherchierbar und darüber hinaus via http://hdl.handle.net/10932/00-0379-FDFE-CC30-0301-E (letzter Zugriff: 8.11. 2017) als downloadbare Ressource über das IDS-Repository erhältlich. Im BBAW-Repository steht das Korpus unter http://hdl.handle.net/11858/00-203Z-0000-002D-EC85-5 (letzter Zugriff: 8.11. 2017) zur Verfügung. Nach kostenfreier Registrierung können Interessierte das Korpus im Portal www.dwds.de im Bereich „Textkorpora" abfragen.

– *DWDS-Blogkorpus*: 103 Millionen Tokens aus CC-lizenzierten Weblogs, großenteils in deutscher Sprache, die über das DWDS-Portal[14] der BBAW recherchierbar sind (Barbaresi & Würzner 2014; Barbaresi 2016).
– *Gießener Scienceblog-Korpus*: laufendes Projekt zum Aufbau eines Korpus deutschsprachiger Wissenschaftler-Blogs an der Universität Gießen (Grumt Suárez, Karlova-Bourbonus & Lobin 2016).
– *Janes*: Projekt *Jezikoslovna analiza nestandardne slovenščine* (*Corpus of Nonstandard Slovene*, Fišer, Erjavec & Ljubešić 2016, 2017)[15] mit 200 Millionen Tokens aus verschiedenen IBK-Formen (Tweets, Forendiskussionen, Blogs, Leserkommentare aus Nachrichtenportalen, Wikipedia-Diskussionsseiten). Die Korpusdaten sind linguistisch annotiert und wurden automatisch nach dem Grad ihrer Konformität zum geschriebenen Standard klassifiziert (Ljubešić et al. 2015); die dabei ermittelten Werte sind den Daten in Form von Metadaten beigegeben.
– *MoCoDa²*: laufendes Projekt an der Universität Duisburg-Essen, in dem eine Datenbank und ein Web-Frontend für die wiederholte Sammlung von Spenden aus digitaler Kurznachrichtenkommunikation (WhatsApp und vergleichbare Dienste) entwickelt wird.[16] Die Datenstücke der Sammlung werden unter Einbeziehung der Spender mit reichhaltigen Metadaten ausgestattet und dadurch insbesondere auch für qualitative korpusgestützte Analysen interessant. Es ist geplant, die aufbereiteten Datenspenden in regelmäßigen Abständen in die Korpussammlungen am IDS zu integrieren.
– *NPS Chat Corpus*: ein Korpus mit 45.000 laufenden Wortformen aus englischsprachigen, altersspezifischen Chatrooms, das um *Part-of-speech*-Informationen und eine Dialogaktklassifikation angereichert ist (Forsyth & Martell 2007) und das über das *Linguistic Data Consortium* (LDC) zur Verfügung gestellt wird.[17]
– *sms4science.ch*: Korpus mit gespendeten SMS-Nachrichten (Deutsch, Französisch, Schweizerdeutsch, Italienisch, Rätoromanisch) im Umfang von 650.000 Tokens (Dürscheid & Stark 2011); online abfragbar in einer Volltextversion (SMS Navigator) und als teilweise annotierte Version in ANNIS.[18]
– *SoNaR-CMC*: IBK-Komponente mit Chats, Tweets und SMS-Nachrichten im Referenzkorpus des Gegenwarts-Niederländischen (Oostdijk et al. 2013); online abfragbar via CLARIN-NL (OpenSoNaR).[19]

---

14 https://www.dwds.de (letzter Zugriff: 8. 11. 2017).
15 http://nl.ijs.si/janes/ (letzter Zugriff: 8. 11. 2017).
16 http://www.mocoda2.de/ (letzter Zugriff: 8. 11. 2017).
17 http://faculty.nps.edu/cmartell/NPSChat.htm (letzter Zugriff: 8. 11. 2017).
18 http://www.sms4science.ch (letzter Zugriff: 8. 11. 2017).
19 https://portal.clarin.nl/node/4195 (letzter Zugriff: 8. 11. 2017).

- *Suomi24*: umfangreiche Sammlung (2,38 Milliarden Tokens) mit Daten aus finnischen Diskussionsforen mit morphosyntaktischen Annotationen; als Download verfügbar.[20]
- *Web2Corpus_it*: laufendes Projekt zum Aufbau eines ausgewogenen Korpus zur italienischen IBK mit Daten aus Foren, Blogs, Newsgroups, sozialen Netzwerken und Chats (Chiari & Calzonetti 2014).[21]
- *whatsup-switzerland.ch*: Korpus des Projekts *What's up, Switzerland?*, in dem spendenbasiert 650 WhatsApp-Chatverläufe im Umfang von insgesamt 5 Millionen Tokens gesammelt wurden.[22]

Der Austausch von *best practices*, Werkzeugen und Expertise zwischen Korpusprojekten zum Thema stellt eine wichtige Vorbedingung dar, um in einem *community-driven approach* projektübergreifend nutzbare, dokumentierte und nachhaltige Lösungen für die Herausforderung zu entwickeln, internetbasierte Kommunikation in linguistischen Korpora zu repräsentieren. Als fruchtbares Format, um die dafür erforderliche Vernetzung anzustoßen, hat sich die Konferenzreihe „Conference on CMC and Social Media Corpora in the Humanities" erwiesen, bei der seit 2013 in jährlichem Turnus Korpusprojekte, Entwickler von korpus- und sprachtechnologischen Werkzeugen sowie korpuslinguistisch arbeitende Geisteswissenschaftler zusammenkommen, um aktuelle Forschungs- und Entwicklungsarbeiten für verschiedene Sprachen vorzustellen sowie Erfahrungen und Lösungsansätze zu diskutieren.[23] Die enge Anbindung an Initiativen und Verbundprojekte, die sich im Bereich der Digital Humanities mit der Etablierung von einheitlichen Formaten (TEI) und Sprachressourceninfrastrukturen (CLARIN, DARIAH) beschäftigen, stellt die Voraussetzung dar, um die Interoperabilität von IBK-Korpora mit Sprachressourcen zu anderen Domänen des Sprachgebrauchs (Text- und Gesprächskorpora) zu gewährleisten.

Interoperabilität: Korpora, die in aufeinander abbildbaren Formaten repräsentiert sind, können einfacher miteinander kombiniert und mit den gleichen Korpustechnologien ausgewertet werden als wenn das nicht der Fall ist. Sind diese Formate zudem kompatibel mit Formaten, die in existierenden Text- und

20 http://urn.fi/urn:nbn:fi:lb-201412171 (letzter Zugriff: 8. 11. 2017).

21 Project page: http://www.glottoweb.org/web2corpus/ (letzter Zugriff: 8. 11. 2017).

22 http://www.whatsup-switzerland.ch/ (letzter Zugriff: 8. 11. 2017). Eine vergleichbare, einmalige Datensammlung für das Deutsche (http://www.whatsup-deutschland.de/, letzter Zugriff: 8. 11. 2017) wurde 2014/2015 von Beat Siebenhaar (Leipzig) initiiert und unter Beteiligung von sieben deutschen Universitätsstandorten durchgeführt.

23 Einen Überblick über Korpusprojekte und darauf bezogene Forschungsfragen bieten die Buchpublikationen zu den Konferenzen 2013, 2015 und 2016 in Dortmund, Rennes und Ljubljana (Beißwenger et al. 2014; Wigham & Ledegen 2017; Fišer & Beißwenger 2017).

Gesprächskorpora verwendet sind, können IBK-Korpora darüber hinaus auch mit Korpora anderen Typs kombiniert und in existierende Korpussammlungen integriert werden. Für die empirische Analyse von Sprache und Interaktion eröffnet das interessante Perspektiven:

1. verbesserte Möglichkeiten der kombinierten Auswertung von IBK-Korpora in unterschiedlichen Sprachen im Rahmen sprachvergleichender Untersuchungen;
2. verbesserte Möglichkeiten der Untersuchung von sprachlichen und kommunikativen Praktiken in unterschiedlichen Formen und Gattungen internetbasierter Kommunikation, die in Korpora unterschiedlicher Anbieter dokumentiert sind;
3. verbesserte Möglichkeiten der kombinierten Auswertung von IBK-Korpora mit Text- und Gesprächskorpora (vergleichende Untersuchung sprachlicher und interaktionaler Phänomene in monologischen Texten, mündlichen Gesprächen und in dialogischer Schriftlichkeit).

Ein wichtiger Baustein für die Herstellung von Interoperabilität ist die Verwendung eines einheitlichen Basisformats für die Repräsentation und Strukturbeschreibung der Korpusdaten. Ein solches Format existiert bislang nicht, es wird aber dringend benötigt.

3 Panta rhei? Internetbasierte Kommunikation zwischen Veränderlichkeit und Stabilität

Die Entwicklung eines Basisformats für die Repräsentation von Daten internetbasierter Kommunikation setzt voraus, dass es übergreifende und stabile Interaktionsformate gibt, die sinnvoll in einem einheitlichen Repräsentationsformat abgebildet werden können. Ein Repräsentationsformat, das für möglichst viele Korpora als grundlegendes Annotations- und als Austauschformat brauchbar sein soll, darf weder so spezifisch sein, dass für jedes Genre internetbasierter Kommunikation (z. B. Chat, SMS, Instant Messaging, Microblogging, Forum, Blogkommentare) bzw. für Daten aus unterschiedlichen Plattformen und Anwendungen (z. B. WhatsApp, Threema, Facebook, Instagram, Twitter, Wikipedia-Diskussionen) ein eigenes Strukturmodell benötigt wird. Stattdessen sollte es einen Abstraktionsgrad aufweisen, der zentrale Strukturmerkmale internetbasierter Kommunikation übergreifend zu Kommunikationsformen und -anwendungen erfasst. Das mit ihm beschriebene Interaktionsformat sollte zudem so stabil sein, dass mit hoher Wahrscheinlichkeit davon ausgegangen

werden kann, dass das Strukturmodell nicht Jahr für Jahr einer Revision bedarf, um es an zwischenzeitlich eingetretene Veränderungen der Technologie anzupassen. Ein Repräsentationsformat, das sich ständig verändert, gewährleistet gerade keine Interoperabilität.

Nun ist jedoch insbesondere die internetbasierte Kommunikation in hohem Maße von Veränderlichkeit geprägt. Das Nachdenken über ein Strukturmodell muss daher zunächst die folgenden Fragen klären:

– Gibt es überhaupt übergreifende Interaktionsformate in der internetbasierten Kommunikation?
– Wenn ja: Sind diese stabil, d. h. relativ beständig gegenüber technologischem Wandel?

Unter einem *Interaktionsformat* verstehe ich dabei eine Struktur für die Organisation von Interaktionsereignissen, die durch kommunikative Rahmenbedingungen geprägt ist, die sich aus Festlegungen auf der Ebene der Kommunikationstechnologie ergeben und für die Nutzer der Technologie nicht änderbar sind. Die Gesamtheit dieser Bedingungen definiert die medialen und modalen Ressourcen, die für die Interaktions*praxis* auf Grundlage der betreffenden Technologie zur Verfügung stehen. Da sich das Format nicht als solches, sondern immer erst bei seiner Instanziierung in konkreten Interaktionsereignissen zeigt, stellt die Konkretisierung des Formats im Vollzug immer die Schnittstelle zwischen den invarianten technologischen Rahmenbedingungen und den sprachlichen und kommunikativen Praktiken dar, mit denen sich die Interagierenden unter Nutzung der gegebenen Ressourcen die technologischen Bedingungen für die Zwecke der Interaktionskonstitution zu eigen machen.

3.1 Veränderlichkeit

Im Kontext der Modellierung und Analyse sprachlicher und kommunikativer Praktiken (Deppermann et al. 2016) bildet die internetbasierte Kommunikation einen spannenden Untersuchungsgegenstand. Geradezu *Fishbowl*-artig lässt sich an (und in) ihr studieren, wie sich die Nutzer internetbasierter Kommunikationstechnologien an die von der Technologie gesetzten Rahmenbedingungen anpassen, um unter diesen Bedingungen bestmöglich das zu tun, wozu die Technologie gemacht ist: Interaktion zu gestalten. Da jedweder Austausch im Netz – das schließt browserbasierte Anwendungen und auf dem Smartphone genutzte, App-basierte Anwendungen gleichermaßen ein – nur unter den Bedingungen technischer Vermittlung zustande kommt und die technischen Rahmenbedingungen für die Interaktionsorganisation in jeder Anwendung unterschiedlich ausgeprägt sein können, bildet die Analyse internetbasierter

Kommunikation ein äußerst produktives Anwendungsfeld, um die Vielfalt und Flexibilität der Praktiken-Praxis unter Bedingungen technischer Vermittlung zu untersuchen. Nicht von ungefähr nimmt die Analyse von Praktiken breiten Raum in der interaktions- und konversationsanalytischen Beschäftigung mit internetbasierter Kommunikation ein. Stark beforscht wurden in der IBK-Forschung schon früh beispielsweise Ausprägungen der Konzepte *turn*, *turn-taking*, *adjacency* und *floor* (Murray 1989; Garcia & Baker Jacobs 1998, 1999; Cherny 1999; Storrer 2001; Beißwenger 2003 u. a.), Praktiken der Kohärenzsicherung (Herring 1999; Severinson Eklundh 2010), die Beitragsproduktion bzw. Turnkonstruktion (Garcia & Baker Jacobs 1998, 1999; Markman 2006; Beißwenger 2007), Formen der Reparatur und der schriftlichen Revision (Schönfeldt & Golato 2003; Beißwenger 2010). Neuere Arbeiten, die Phänomene internetbasierter Kommunikation auf dem Hintergrund des Programms der Interaktionalen Linguistik (Selting & Couper-Kuhlen 2000) analysieren, sind z. B. Imo (2015a) zur konstruktionsgrammatischen Analyse von Emoticons, Imo (2015b) zur Portionierung von Äußerungen in WhatsApp-Dialogen, König (2015) zu Sequenzmustern, Lindemann, Ruoss & Weinzinger (2014) zu Praktiken des Editierens. Einen Überblick über Kandidaten für Praktiken in verschiedenen Formen internetbasierter Kommunikation bietet Beißwenger (2016).

Praktiken variieren in Abhängigkeit zu den modalen und medialen Ressourcen, die den Interaktionsbeteiligten für die Interaktionskonstitution zur Verfügung stehen. Die Ressourcen, die in Formen der technisch vermittelten Kommunikation genutzt werden können, variieren in Abhängigkeit zur Veränderlichkeit von Technologien. In technologischer Hinsicht ist Veränderlichkeit ein fundamentales Merkmal der internetbasierten Kommunikation: Neue Kommunikationstechnologien kommen auf und verdrängen andere Kommunikationstechnologien aus der Gunst der Nutzer bzw. bestimmter Nutzergruppen. So hat beispielsweise die Anwendung Snapchat nach den Ergebnissen der jüngsten Ausgabe der JIM-Studie (2016) in der Gunst der 12–19-Jährigen deutlich zugelegt, während für Facebook bei derselben Gruppe ein vergleichbar großer Abfall in der Nutzergunst zu verzeichnen ist. Zugleich entwickeln sich Kommunikationstechnologien weiter, so dass sich für die Nutzer der auf ihrer Grundlage bereitgestellten Anwendungen die zur Verfügung stehenden technischen Handlungsmöglichkeiten und damit die medialen und modalen Ressourcen für die Interaktionskonstitution kontinuierlich verändern. So hat die Plattform Facebook 2013 die ursprünglich nur auf Twitter nutzbaren Hashtags eingeführt, um ihren Nutzern zusätzliche Möglichkeiten zur thematischen Vernetzung von Beiträgen zu bieten. Durch diese Änderung hat sich das System von Mitteln für die thematische Organisation und Verknüpfung von Interaktionen insgesamt verändert. Ähnliches gilt für die – ebenfalls von Twitter

übernommene – Einführung der automatischen Verlinkung von Adressierungsausdrücken auf die Profilseite des adressierten Nutzers, verbunden mit der Anzeige des Postings, das eine Adressierung enthält, auf der individuellen Startseite des Adressaten: Adressierung wird dadurch in Facebook zu einem Mittel, um die Aufmerksamkeit von Adressaten zu gewinnen und diesen die Existenz für sie relevanter Interaktionsäußerungen im *Push*-Verfahren bekannt zu machen. Damit gewinnt die Praktik des Adressierens, wie sie sich in der frühen Chat- und Foren-Kommunikation als zunächst rein textuelle Praktik für die Kohärenzsicherung herausgebildet hat, an zusätzlichen Dimensionen und entwickelt sich zu einer Praktik der Aufmerksamkeitsgewinnung.

3.2 Stabilität

Das Internet ist mittlerweile alt genug für Stabilität. Bei all der Variabilität von Ressourcen und darauf bezogenen Praktiken gibt es durchaus Formate, die sich als weitgehend beständig gegenüber Wandel erwiesen haben und die sich, auch wenn sie in unterschiedlichen Instanziierungen unterschiedlich ausgeprägt sein mögen, übergreifend zu einzelnen IBK-Genres und -Plattformen beschreiben lassen. Im Folgenden charakterisiere ich anhand einer Gegenüberstellung der Instanziierungen von Interaktion in sechs verschiedenen Arten von Kommunikationsumgebungen ein Format, das seit inzwischen mehr als einem Vierteljahrhundert in der internetbasierten Kommunikation präsent ist und das sicherlich auch noch in den nächsten zehn Jahren Bestand haben wird. Zwar wandelt auch dieses Format kontinuierlich sein Gesicht und stellt sich in unterschiedlichen Instanziierungen durchaus unterschiedlich dar; im Kern hat es sich jedoch als erstaunlich robust erwiesen.

Zwischen den in den Abb. 14.1–14.6 veranschaulichten Instanziierungen von Interaktion gibt es Gemeinsamkeiten und Unterschiede. Gemeinsam ist ihnen, dass sie aus schriftlichen Einheiten aufgebaut sind, die im Layout klar voneinander abgegrenzt sind und die jeweils einem Urheber zugeordnet sind. Ich nenne diese Einheiten *Postings*, weil das englische Verb ‚to post' und sein ins Deutsche entlehntes Pendant ‚posten' treffend ausdrücken, was diese Einheiten charakterisiert und von den grundlegenden Einheiten mündlich realisierter Interaktionen – den Turns – unterscheidet:

– Sie werden erst *als Ganze* für die anderen Beteiligten wahrnehmbar, und zwar erst dann, nachdem sie durch Ausführung einer expliziten Verschickungsanweisung an den Server übergeben und von diesem weiterübermittelt wurden. Eine inkrementelle Verarbeitung der Äußerung (= des Geposteten) simultan zu ihrer Hervorbringung und damit eine rezipienten-



Abb. 14.1: Vier Postings mit Thread-Struktur auf einer Facebook-Profilseite: Die Postings 3 und 4 erscheinen, da sie von den Verfassern als „Antworten" auf das Vorgänger-Posting gekennzeichnet wurden, eingerückt.

seitige Einwirkung auf den Verbalisierungsplan des Produzenten, wie das für Turns in mündlichen Gesprächen charakteristisch ist, ist ausgeschlossen.

– Die Produktion geht der Wahrnehmung durch andere voraus und ist ein privater Akt, auch wenn manche Kommunikationsanwendungen – z. B. WhatsApp – ihren Nutzern über temporäre Hinweise anzeigen, wenn einer der anderen Beteiligten gerade einen Beitrag eingibt („Nutzer X schreibt ..."). Was während der Beitragseingabe im Eingabeformular auf dem Display des Produzenten entsteht und wie es sich verändert, bleibt den übrigen Beteiligten aber verborgen.
– Solange der eingegebene Beitrag nicht verschickt wurde, hat der schriftliche Entwurf den Charakter der Vorläufigkeit – und zwar auch technisch: Er kann ganz oder in Teilen, beliebig oft und in beliebigem Umfang revidiert werden. Beißwenger (2007, 2010) zeigt empirisch an dokumentierten Produktionsverläufen, dass von dieser Möglichkeit rege Gebrauch gemacht wird. Eine

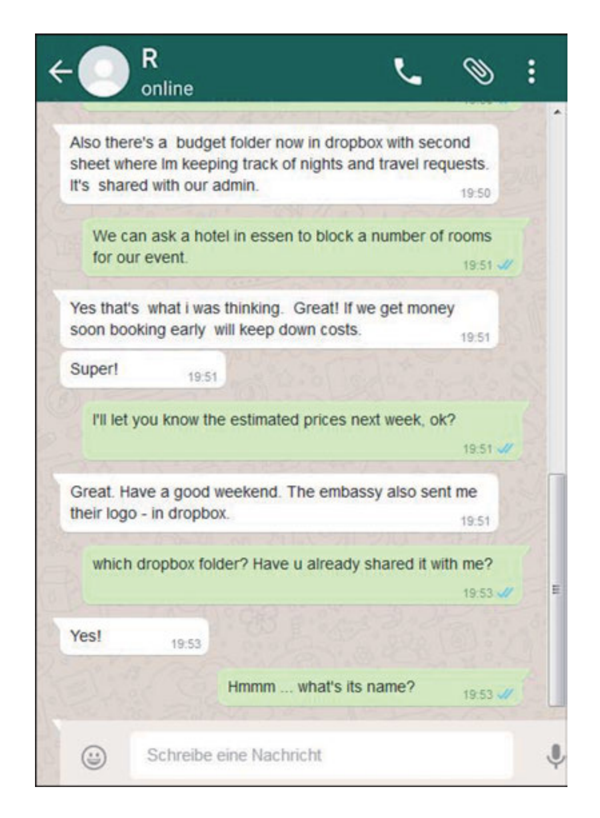


Abb. 14.2: Ausschnitt aus einer WhatsApp-Interaktion. Die rechtsbündig dargestellten Postings wurden von der Interaktionspartnerin verfasst und sind mit automatisch generierten Hinweisen zum Übermittlungsstatus (Häkchen rechts unten) versehen.

Bearbeitung des Entwurfs ist grundsätzlich nicht nur für den Autor selbst, sondern auch für Zusatzprogramme möglich, die auf dem Endgerät oder in der betreffenden Kommunikationsanwendung im Hintergrund laufen (z. B. Autokorrekturfunktionen, die, insbesondere auf mobilen Endgeräten, z. T. erheblich zur sprachlichen Gestalt der verschickten Postings beitragen).

Einen Sonderfall der Postings stellen die sogenannten Sprachnachrichten in WhatsApp-Interaktionen dar. Dabei handelt es sich um aufgezeichnete, mündlich realisierte Äußerungen, die in Form von Audiodateien übermittelt und dadurch – analog zu ihrem schriftlichen Pendant – erst im Nachhinein zur Verbalisierung rezipierbar werden. Ich bezeichne diese Form der Postings als *Audio-Postings*, um sie – aus denselben Gründen wie im Falle der schriftlichen Postings – terminologisch von Turns in Gesprächen abzugrenzen. Revisionen

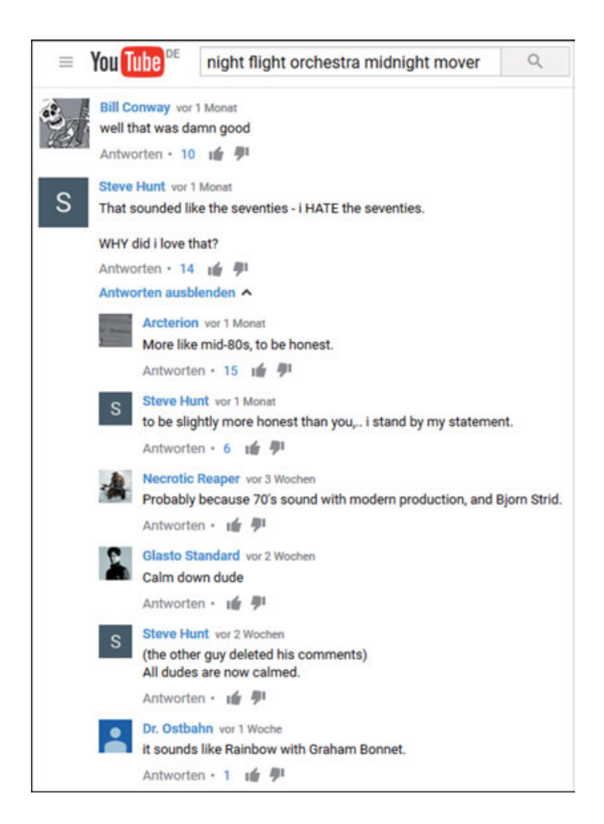


Abb. 14.3: Kommentarsektion unter einem YouTube-Video. Die einzelnen Postings sind – ähnlich wie in Facebook – mit sog. „Likes" versehen; von den Verfassern als Antworten markierte Postings erscheinen eingerückt.

von Sprachnachrichten sind nur komplett möglich: Eine laufende Audioaufzeichnung kann abgebrochen, aber nicht in Teilen editiert werden. Für übermittelte Audio-Postings gilt, dass der Prozess ihrer Verbalisierung in der Rezeption für die Adressaten transparent wird; das gilt aber nicht für Vorversionen, bei denen die Aufnahme abgebrochen wurde. Auch das ist ein wichtiger Unterschied zu den Verbalisierungs- und Wahrnehmungsbedingungen in mündlichen Gesprächen.

Gemeinsam ist den abgebildeten Beispielen weiterhin Folgendes:

– Die Vermittlung der Interaktionsbeiträge erfolgt über ein *Logfile* am Bildschirm, das im Falle „synchroner"[24] Kommunikationsumgebungen (z. B. in

24 Ich verwende den Ausdruck ‚synchron' in Anführungszeichen, um deutlich zu machen, dass Chats aufgrund der oben beschriebenen Zeitlichkeitsbedingungen nicht in gleicher Weise unter Bedingungen der „Gleichzeitigkeit" stattfinden wie mündliche Gespräche. Auf Garcia &



Abb. 14.4: Posting auf Twitter („Tweet“) mit zwei Antwort-Postings, die als Thread unter dem Bezugsposting angezeigt werden. Die einzelnen Postings enthalten automatisch generierte Metadaten (u. a. zur Anzahl der Likes und Antworten sowie zur Anzahl der Retweets).

klassischen IRC- und Webchats) mindestens für die Dauer der aktuellen Sitzung, in vielen Umgebungen aber theoretisch unbegrenzt persistent ist (WhatsApp, Online-Foren, Twitter, Facebook usw.). Was geäußert wird, dokumentiert sich in einem gespeicherten Verlauf. Durch das Posting-Format der ausgetauschten Beiträge wird die Persistenz von Interaktionsbeiträgen zur notwendigen Grundbedingung für die Interaktionskonstitution: Wenn Beiträge zunächst als Ganze verbalisiert werden, bevor sie in einem der Verbalisierung nachgeordneten Schritt für die Adressaten wahrnehmbar werden, bedarf die Interaktion der *Überlieferung* (i. S. v. Ehlich 1984), um zu funktionieren. Da überdies der Zeitpunkt, ab dem ein Posting am Bild-

---

Baker Jacobs (1998) geht der Vorschlag zurück, die Kommunikation in Chats und mit vergleichbaren Technologien als *quasi-synchronous* (bzw. mit Dürscheid 2005 als „quasi-synchron“) zu charakterisieren. Das Etikett ‚quasi-synchron‘ signalisiert zwar, dass die Kommunikation in den damit charakterisierten Form(at)en nicht gänzlich synchron verläuft, bleibt hinsichtlich des Unterschieds, um den es geht, jedoch unterspezifiziert bzw. weist ihn als letztlich doch vernachlässigbar aus (lat. *quasi* = ‚vergleichbar, ungefähr‘). Gerade im ‚quasi‘ zeigen sich allerdings die für die Interaktionskonstitution fundamentalen Unterschiede zum Gespräch. In Beißwenger (2007, 2010) habe ich deshalb vorgeschlagen, Formen wie Chats als „synchrone, aber nicht simultane“ Formen bzw. die Kommunikationsbedingungen in Chats und vergleichbaren

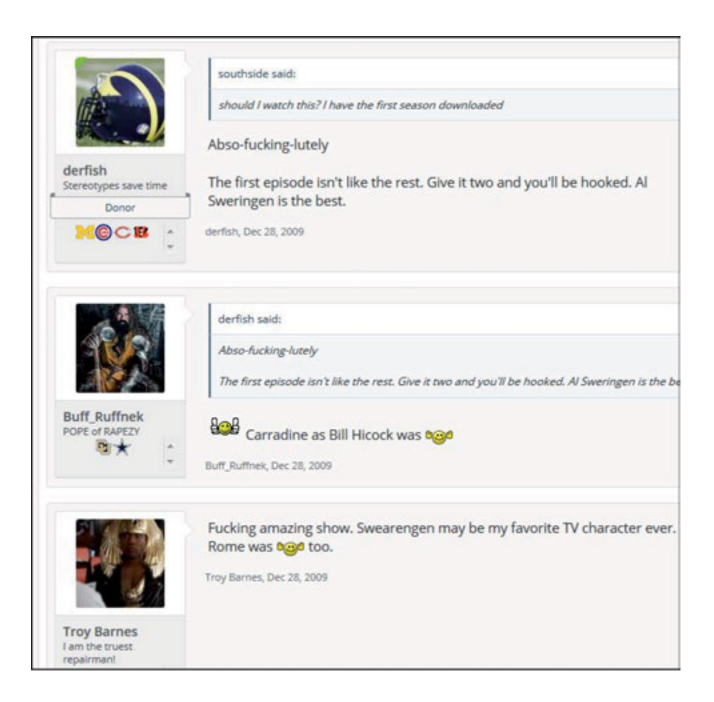


Abb. 14.5: Ausschnitt aus einem Thread mit einem Online-Forum. Die Verfasser der Postings 1 und 2 nutzen die Möglichkeit, vorangegangene Postings anderer Urheber ganz oder in Teilen zu zitieren. Die Zitate werden vom System automatisch in den Beiträgen reproduziert; dabei werden zudem automatisch generierte Sprachbausteine eingefügt („southside said", „derfish said"). Auch die einzelnen Postings nebengestellter Ausschnitte aus den hinterlegten Nutzerprofilen der Verfasser wurden vom System automatisch reproduziert.

schirm wahrnehmbar ist, nicht notwendigerweise mit dem Zeitpunkt seiner tatsächlichen Wahrnehmung durch die Adressaten zusammenfällt, sichert das *Logging* der Postings nicht nur die Bearbeitung der zeitlichen Zerdehnung zwischen Produktion und Zustellung, sondern zudem die Bearbeitung der Zerdehnung zwischen Zustellung und tatsächlicher Rezeption. Dass selbst in „synchronen" Formen wie dem klassischen Chat, in

Formen als „synchron ohne Synchronisierung" zu beschreiben. Synchronizität wird durch diesen Vorschlag in zwei Aspekte zerlegt: (i) den Aspekt der zeitgleichen Orientiertheit der Beteiligten auf die Entwicklung des Kommunikationsgeschehens, die auch der vorwissenschaftlichen und alltagssprachlichen Charakterisierung von Chats als „synchronen" Formen zugrunde liegt; (ii) den Aspekt der Alignierung von Verbalisierung, Übermittlung und Verarbeitung, die im Falle mündlicher Gespräche vollumfänglich gegeben (i. S. v. der von Auer 2000 beschriebenen ‚On-line'-Verbalisierung), im Falle von Chats aber durch eine Zerlegung in eine konsekutive Abfolge von Schritten gekennzeichnet ist (Beißwenger 2007: 35–37; Beißwenger 2010: 257–258).

Offrande au Saint Sacrement [Quelltext bearbeiten]

Dieses Orgelstück kenne ich nicht, da es offensichtlich in keiner Messiaen-Orgel-Gesamtaufnahme enthalten ist. Wie ist das zu erklären? Und welcher Art ist das Stück (groß, klein; mit gregorianischem Thema?) Humpyard 00:34, 28. Jan. 2008 (CET)

Das weiss ich natürlich auch nicht. Eine Möglichkeit wäre, es handelt sich um Unkenntnis. Vielleicht sind manchmal Titel nicht eindeutig? Vielleicht wird etwas als "Orgel-Gesamtaufnahme" bezeichnet, wenn die Aufnehmer vermuten, es sei gesamt? Letzte Idee: Vielleicht hat es der Komponist nicht ausdrücklich für Orgel ausgewiesen.- All dies sind nur theoretisierende Denkansätze! --888344

Oder auch - kucke mal hier: http://www.joergabbing.de/publikationen.htm --888344

in der Latry-Gesamtaufnahme (DGG) ist das Stück enthalten (CD2), die Google-Suche fördert außerdem mehrere Quellen zu Tage, die Partitur des Werkes ist z.B. bei Leduc erschienen Gruß Akeuk 20:02, 2. Feb. 2008 (CET)

Abb. 14.6: Ausschnitt aus einer Wikipedia-Diskussionsseite. Der Grad der Einrückung der einzelnen Postings wurde von den Verfassern bei der Eingabe manuell erzeugt.

denen die Interaktionsbeteiligten zeitgleich auf die Teilhabe an der Weiterentwicklung des Interaktionsgeschehens orientiert sind, Beiträge häufig erst im Nachhinein zu ihrer Verfügbarkeit am Bildschirm gelesen werden, zeigt Beißwenger (2007).

– Im Logfile, das als Rezeptionsgrundlage fungiert, erscheinen die Postings in einer klaren Abfolge. Diese muss nicht mit der von den Interaktionsbeteiligten intendierten Handlungsabfolge identisch sein. Die Zerdehnung von Produktion und Übermittlung begünstigt, insbesondere in „synchronen“ Gruppeninteraktionen, die Überkreuzung von Posting-Sequenzen, die unterschiedliche Handlungssequenzen realisieren (Herring 1999; Storrer 2001). In „asynchronen“ Formen nutzen die Schreiber Möglichkeiten des Threadings und der Arbeit mit Zitaten, um die sequenzielle Einordnung ihrer Beiträge anzuzeigen und nachvollziehbar an bestimmte Positionen der Sequenz anzuknüpfen (z. B. Severinson-Eklundh 2010).
– Das *Logging* des Kommunikationsverlaufs am Bildschirm schafft gegenüber mündlichen Interaktionen veränderte Bedingungen für den Bezug auf frühere Teile der Sequenz: Postings können mehrfach und wiederholt rezipiert werden. Das gilt für medial schriftlich realisierte Postings in gleicher Weise wie für Postings, die in Form aufgezeichneter Audioaufnahmen realisiert sind (Sprachnachrichten in WhatsApp). Wann ein Posting interaktional bearbeitet und beantwortet wird, kann flexibler gehandhabt werden als im mündlichen Gespräch. In Online-Foren, die über viele Jahre

laufen, lässt sich beobachten, dass bisweilen an Postings oder Threads angeknüpft wird, die mehrere Monate oder gar Jahre alt sind. Und selbst in Anwendungen wie WhatsApp, die eine „synchrone" Nutzung erlauben, werden Postings nicht immer unmittelbar beantwortet – auch dann nicht, wenn sie von den Adressaten unmittelbar nach ihrer Zustellung rezipiert wurden. Die Persistenz des Verlaufs erlaubt es vielmehr, die Teilhabe an Interaktionen flexibel an das individuelle Zeit- und Aufgabenmanagement anzupassen.

– Der Interaktionsverlauf ist digital gespeichert und mit Standardwerkzeugen von Betriebssystemen verwalt- und bearbeitbar: Eigene und fremde Postings sowie Teile davon (d. h. auch die darin integrierten Bild-, Ton- und Videodateien) können in die Zwischenablage kopiert und in anderen Kontexten reproduziert werden; die am Bildschirm vorgehaltenen Verläufe können in Dateien gespeichert und weiterverwendet werden.
– Neben dem vom Verfasser eingegebenen sprachlichen Inhalt (*user generated content*) können Postings automatisch, d. h. vom System generierte oder reproduzierte, Beitragsteile enthalten. Dazu zählen beispielsweise die automatisch eingefügten Zitate von Postings oder Posting-Teilen anderer Schreiber in Abbildung 14.5, im weiteren Sinne aber auch die vom System hinzugefügten, im Layout den Beiträgen zugeordneten und textuell und/ oder in Form von Grafiken repräsentierten Metadaten zu den Postings und ihren Produzenten in den Beispielen 14.1–14.6.

Daneben weisen einzelne der in den Abbildungen 14.1–14.6 gezeigten Beispiele auch Merkmale auf, die nicht generell, sondern nur spezifisch für einzelne Kommunikationsumgebungen gelten. Diese Merkmale sollen hier nicht weiter thematisiert werden, da bei der Entwicklung eines Basisschemas zur Repräsentation von Strukturmerkmalen zunächst diejenigen Merkmale im Vordergrund stehen, die übergreifend zu einzelnen Kommunikationsumgebungen eine Rolle spielen. Das schließt nicht aus, dass auch weitere, spezifischere Merkmale in ihm abgebildet werden können oder dass das Schema in einem konkreten Projekt in Hinblick auf die speziellen Erfordernisse der technischen Instanziierung von Interaktion in einzelnen IBK-Umgebungen erweitert werden kann. Der Fokus liegt im Folgenden auf der texttechnologischen Modellierung des Gemeinsamen und seiner Darstellung in Form einer Extension zu den *Guidelines for Text Encoding* der TEI.

4 Ein Basisschema für die Repräsentation internetbasierter Kommunikation in TEI

Ein Basisschema, das für möglichst viele Projekte sinnvoll ist, eine Interoperabilität mit bestehenden Sprachressourcen gewährleistet und flexibel einsetzbar ist, sollte die folgenden Merkmale aufweisen:

1. Es sollte auf etablierte Repräsentationsstandards für Sprachressourcen im Bereich der Korpusliguistik und der Digital Humanities bezogen sein.
2. Es sollte sich für die Bedürfnisse konkreter Korpusprojekte und Forschungsfragen erweitern lassen, ohne die Bedingung (1) zu verletzen.
3. Es sollte so allgemein gehalten sein, dass es als Grundlage für die Annotation in vielen Korpusprojekten nützlich ist.
4. Es sollte so spezifisch sein, dass es grundlegende strukturelle Besonderheiten von IBK-Daten – insbesondere die charakteristischen Unterschiede zur Struktur von redigierten Texten und von Gesprächsverläufen – erfasst.
5. Es sollte in seinen obligatorischen Strukturelementen weitestgehend automatisiert aus gesammelten Rohdaten erzeugt werden können.
6. Es sollte die Erzeugung anonymisierter Sichten auf die Korpusdaten unterstützen.

Anforderung (1) wird im nachfolgend vorgestellten Schema eingelöst durch die Orientierung an den Richtlinien der TEI, die seit 1987 Formate für die Strukturbeschreibung textueller Sprachdaten in den Geisteswissenschaften entwickelt. Die Vorschläge der TEI, die seit 1994 in Form von *Guidelines* vorliegen, können als ein *De-facto*-Standard im Bereich der Geisteswissenschaften gelten, der einer Vielzahl digitaler Forschungsressourcen zugrunde liegt. Der Standard wird von einer breiten Nutzercommunity gepflegt und kontinuierlich weiterentwickelt. Über die Konsistenz der Guidelines wacht ein *Technical Council*. Grundsätzlich kann jeder Interessierte Eingaben für die Weiterentwicklung des Standards machen. Zu Themen, denen die TEI-Community in Hinblick auf die Pflege und Anpassung der Guidelines besondere Relevanz beimisst, sind *Special Interest Groups* (kurz: SIGs) eingerichtet, die sich mit der Nutzung von TEI in einzelnen Anwendungsbereichen oder für bestimmte Textgenres und Kommunikationsbereiche befassen (z. B. *TEI for Linguists*, *Correspondence*, *Manuscripts*, *Scholarly Publishing*).[25] Seit 2013 gibt es eine SIG *Computer-mediated*

25 Die gegenwärtig aktiven SIGs sind auf der Seite http://www.tei-c.org/Activities/SIG/ (letzter Zugriff: 8. 11. 2017) verzeichnet.

Communication (TEI-CMC-SIG), die sich mit der Entwicklung TEI-konformer Formate für Sprachressourcen internetbasierter Kommunikation befasst.

Anforderung (2) wird dadurch gewährleistet, dass TEI die schon erwähnte Möglichkeit der *customization* vorsieht (vgl. Abschnitt 1). Diese erlaubt es, aus dem in den TEI-Richtlinien vorgesehenen Inventar an Modellen (Elementklassen, Elementen, Attributen) Auswahlen zu treffen oder bestimmte Teilinventare zu unterdrücken, neue Elemente und Attribute hinzuzufügen oder bestehende Modelle zu ändern. Solange dabei bestimmte Regeln eingehalten werden, die sicherstellen, dass die vorgenommenen Änderungen sich konsistent in die Architektur des TEI-Frameworks einfügen, stimmt das resultierende Schema grundsätzlich mit den TEI-Richtlinien überein. Für die Erzeugung eigener *customizations* bietet die TEI mit dem Editor *Roma* sogar ein spezialisiertes Tool an.[26]

Da die TEI-Richtlinien bislang keine spezifischen Modelle für die Strukturannotation von IBK-Daten anbieten, arbeitet die TEI-CMC-SIG bei der Entwicklung eines Basisschemas mit der Möglichkeit der *customization*. Bis dato haben Mitglieder der SIG drei Annotationsschemas entwickelt und diese jeweils an Korpora getestet. Der Vorteil der Orientierung an einem Standard und dessen Erweiterung (per *customization*) macht die Schemaentwicklung darüber hinaus ökonomisch: Diejenigen Teile des Schemas, die für die Repräsentation von Korpusdaten benötigt werden, aber keine IBK-spezifischen Merkmale abbilden, können aus dem in TEI schon vorhandenen Standardinventar übernommen werden; Anpassungen und Erweiterungen sind nur da erforderlich, wo der TEI-Standard noch keine geeigneten Lösungen anbietet. Entsprechend stimmen die von der SIG entwickelten Schemas in weiten Teilen mit dem TEI-Standard überein; diejenigen Modelle, die tatsächlich neu eingeführt oder gegenüber dem Standard modifiziert werden mussten, bilden nur einen kleinen Teil.

Die Anforderungen (3), (4) und (5) werden dadurch eingelöst, dass sich das Schema auf die Modellierung der Eckpunkte eines Interaktionsformats beschränkt, das vielen Instanziierungen internetbasierter Kommunikation zugrunde liegt, und dass es sich aus Strukturinformation erzeugen lässt, die in den Korpus-Ausgangsdaten i. a. R. bereits vorhanden ist. Die Anforderung (6) wird eingelöst durch eine Transformation (mindestens) solcher personenbezogener Daten, die sich anhand von Strukturmerkmalen in den Ausgangsdaten identifizieren lassen, in Metadaten.

Die TEI-CMC-SIG hat in einem iterativen Prozess bislang drei Schemas für die Strukturannotation von IBK-Korpora entwickelt. Das erste Schema (Beiß-

26 http://www.tei-c.org/Guidelines/Customization/use_roma.xml (letzter Zugriff: 8. 11. 2017).

wenger et al. 2012; „DeRiK-Schema“[27]) wurde in Zusammenarbeit mit dem Team des DWDS-Projekts an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) als Grundlage für die Repräsentation von IBK-Erweiterungen für Korpora geschriebener Sprache (Beißwenger & Lemnitzer 2013) konzipiert. Das zweite Schema (Chanier et al. 2014; „CoMeRe-Schema“[28]) wurde als Ergebnis verschiedener Diskussionen in der neu gegründeten TEI-CMC-SIG und auf der Grundlage des DeRiK-Schemas von Thierry Chanier und Kollegen für die Repräsentation einer Sammlung von 14 IBK-Korpora zum Französischen entwickelt (CoMeRe-Korpora). Die dritte Schemaversion (CLARIN-D-Schema) entstand im Rahmen des CLARIN-D-Kurationsprojekts *ChatCorpus2CLARIN*, in dem das Dortmunder Chat-Korpus in die Korpusinfrastrukturen am IDS und an der BBAW integriert und in diesem Zusammenhang in TEI remodelliert wurde (Lüngen et al. 2016). Dieses dritte Schema basiert wiederum auf einer Analyse des CoMeRe-Schemas und der mit dessen Anwendung gemachten Annotationserfahrungen und baut das Schema für verschiedene Formen schriftbasierter IBK weiter aus.

Im Folgenden beschreibe ich den letzten Stand des Schemas, das CLARIN-D-Schema.[29] Ich stelle einige wesentliche Merkmale vor, die die in Abschnitt 3 dargestellten Charakteristika schriftbasierter IBK im TEI-Kontext und mit Blick auf die oben formulierten Anforderungen umsetzen. Die bei der Entwicklung des Schemas in das TEI-Rahmenwerk eingebrachten Modifikationen und Erweiterungen sind in einem ODD-Dokument (*One Document Does It All*) beschrieben, das unter derselben Adresse in dem von der TEI vorgegebenen Dokumentationsformat angeboten wird.[30]

In Bezug auf die Repräsentation von IBK-Daten unterscheide ich im Folgenden die folgenden Datentypen und Strukturebenen:[31]

– *Primärdaten:* die eigentlichen Nutzdaten, d. h. diejenigen Sprachdaten und ihre Struktur, an deren Auswertung Nutzer linguistischer Korpora interes-

---

27 https://wiki.tei-c.org/index.php/SIG:CMC/derikschema (letzter Zugriff: 8. 11. 2017).

28 https://wiki.tei-c.org/index.php?title=SIG:CMC/comereschema (letzter Zugriff: 8. 11. 12017).

29 Das vollständige Schema steht unter https://wiki.tei-c.org/index.php?title=SIG:CMC/clarindschema (letzter Zugriff: 8. 11. 2017) als Schemaspezifikation in der XML-Schemasprache *Relax NG* zur Verfügung und kann in dieser Form für die Korpusannotation eingesetzt werden. An der Entwicklung des Schemas unmittelbar beteiligt waren neben dem Verfasser Axel Herold, Harald Lüngen, Angelika Storrer und Eric Ehrhardt.

30 Zum ODD-Format vgl. Lobin (2010: 110–111).

31 Für die Unterscheidung von Primärdaten, Metadaten und Annotationen greife ich auf eine in der Korpuslinguistik allgemein übliche Differenzierung zurück (vgl. z. B. Storrer 2011: 2018–2219, Perkuhn, Keibel & Kupietz et al. 2012: 45, Lemnitzer & Zinsmeister 2015: 13), die Unterscheidung zweier Strukturebenen in den Primärdaten folgt Beißwenger et al. (2012).

siert sind. In Bezug auf IBK-Daten geht das vorgestellte Schema davon aus, dass es sich bei den Ausgangsdaten, die die Primärdaten für IBK-Korpora bereitstellen, um Dokumente handelt, die eine Abfolge von Postings in einer bestimmten Art der Strukturierung enthalten, die von zwei oder mehreren Autoren produziert wurden (sogenannte *Logfiles* oder Mitschnitte). Strukturen in IBK-Primärdaten lassen sich dabei auf zwei Ebenen beschreiben:

a) auf der Ebene der *Makrostrukturen*: Darunter verstehe ich die spezifische Art der Abfolge und Anordnung von Postings in den Ausgangsdokumenten und im Bildschirmprotokoll der Beteiligten. Makrostrukturen werden charakteristischerweise nicht von einem Autor alleine hergestellt, sondern ergeben sich aus dem Zusammenspiel von Verschickungshandlungen zweier oder mehrerer Autoren *plus* den Aufbereitungs- und Vermittlungsroutinen der zugrunde liegenden Kommunikationstechnologie.
b) auf der Ebene der *Mikrostrukturen*: Darunter verstehe ich die Struktur von Postings und damit all diejenigen Formen der Strukturierung sprachlicher Äußerungen, die der alleinigen Gestaltungshoheit des Autors des Postings unterliegen. Mikrostrukturen lassen sich mit Blick auf die visuelle Gliederung des Posting-Inhalts und die Integration von Medienobjekten beschreiben (Enthält das Posting nur einen oder mehrere Absätze? Enthält es Zwischenüberschriften? Ist mit Listenformatierungen und Einrückungen gearbeitet? Sind Bilder, Audio- oder Videodateien integriert?); Mikrostrukturen können aber auch unter textgrammatischem und syntaktischem Aspekt von Interesse sein (Welche Formen der Verknüpfung von Sätzen nutzt der Autor? Welche syntaktische Struktur weist sein Posting auf?).

– *Metadaten:* Daten, die die Primärdaten beschreiben und die benötigt werden, um die im Korpus dokumentierte Primärdatenstichprobe als solche und auch die darin dokumentierte Form der Sprachverwendung (die im Falle von IBK-Daten interaktional organisiert ist) zu kontextualisieren.
– *Annotationen:* Daten, mit denen Primärdatensegmenten linguistische oder Strukturinformationen zugeordnet werden und die im nachfolgend beschriebenen Basisschema dazu verwendet werden, Strukturinformation auf Makro- und Mikroebene in Form expliziter Beschreibungen zu repräsentieren.

4.1 Metadaten und TEI-Header

Das Schema folgt in seiner Architektur den obligatorischen Strukturvorgaben für TEI-Schemas und interpretiert diese IBK-spezifisch. Standardmäßig umfasst

jedes TEI-Dokument einen *Header*, in dem Metadaten zum Dokument erfasst werden, gefolgt von der Repräsentation der annotierten Daten. Listing 1 zeigt einen Ausschnitt aus der Header-Struktur für ein Dokument aus dem Dortmunder Chat-Korpus. Die Teilstruktur *fileDescription* (Element `fileDesc`) erfasst im *publicationStatement* (Element `publicationStmt`) Angaben zum Anbieter der Ressource (im vorliegenden Fall das IDS und die BBAW) und zu den Lizenzbedingungen, unter denen die Ressource genutzt werden darf (CC BY 4.0) sowie in der *sourceDescription* (Element `sourceDesc`) Angaben zur Herkunft der Daten (in diesem Fall zur Chat-Umgebung, in der das enthaltene Logfile aufgezeichnet wurde, sowie zu Datum und Uhrzeit der Aufzeichnung).[32]

Listing 1: Ausschnitt 1 (gekürzt) aus dem TEI-Header für ein Dokument aus dem Dortmunder Chat-Korpus in CLARIN-D:

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
    <teiHeader>
      <fileDesc>
        <publicationStmt>
          <publisher/>
          <pubPlace/>
          <idno>1204009</idno>
          <distributor>
           CLARIN centre Institut für Deutsche Sprache, Mannheim
           and CLARIN centre Berlin-Brandenburgische Akademie der
           Wissenschaften
           </distributor>
           <date>2016</date>
           <availability>
             <licence target="https://creativecommons.org/ licen
             ses/by/4.0/">CC BY 4.0. Legal restrictions may arise
             from data protection legislation.</licence>
           </availability>
         </publicationStmt>
         <sourceDesc>
          <recordingStmt>
            <recording>
              <equipment>
```

32 Definition und Inhaltsmodell sämtlicher Elemente aus dem TEI-Standard lassen sich in den *Guidelines* nachschlagen (TEI-P5).

```
          <p>plattformName=<name type="OTH">[_CHATPLATFORM
          NAME_]</name></p>
          <p>plattformURL=<ref type="URL">[_WWWURL_]</ref>
          </p>
        </equipment>
        <respStmt>
          <persName xml:id="f1204009.p1">unknown</persName>
          <resp>recording</resp>
        </respStmt>
        <date>2005-01-11</date>
        <time from="22:07:00" to="22:59:00"/>
      </recording>
    </recordingStmt>
...
</fileDesc>
```

Listing 2 zeigt einen weiteren Ausschnitt aus dem TEI-Header. Die Teilstruktur *profileDescription* (Element `profileDesc`) klassifiziert den Dokument-Inhalt mit Referenz auf ein externes Klassifikationsschema. Die Benennung des Elements `textClass` zeigt, dass TEI ursprünglich für die Annotation redigierter, schriftlicher Texte entwickelt wurde, für die die Zuordnung zu einer Textsorte intendiert ist. Im Falle des Chat-Korpus wurde anstelle eines Textklassifikationsschemas eine Klassifikation der enthaltenen Logfiles nach gesellschaftlichen Handlungsbereichen und auf zweiter Klassifikationsebene nach Chat-Plattformen vorgenommen. Unter der im Listing referenzierten URL ist das für das Korpus verwendete Klassifikationsschema zentral hinterlegt. Durch Aufruf der *target*-URL lässt sich entnehmen, dass das im Dokument beschriebene Logfile der Hauptklasse „Professionelle Chats: Chat-Veranstaltungen im Hochschulkontext, Chat-Beratung, Chats im Medienkontext" sowie innerhalb dieser Klasse der Subklasse „Beratung" zugeordnet ist und dass das Logfile einem Beratungsangebot mit dem Thema „Beratung durch eBay-Expertin" entstammt.

Die Teilstruktur *participationDescription* (Element `particDesc`) beschreibt gemäß Definition „the identifiable speakers, voices, or other participants in any kind of text or other persons named or otherwise referred to in a text, edition, or metadata"; für das Chat-Korpus wird das Element für die Hinterlegung von Informationen zu den Chat-Beteiligten adaptiert. Für jeden Beteiligten gibt es in der *listPerson* einen Eintrag mit eindeutiger ID, dem ein Teilnehmername, eine Teilnehmerrolle (hier: *system*, *expert*, *client*) sowie eine Angabe zum (erschlossenen oder vermuteten) biologischen Geschlecht (*sex*) zugeordnet ist. Da das Dortmunder Chat-Korpus für die Integration in CLARIN-D anonymisiert

wurde, sind die ursprünglichen Teilnehmernamen im Listing durch Kategorienlabels ersetzt (zur Anonymisierung vgl. Beißwenger et al. 2017b). Die ID ist wichtig, um im beschriebenen Logfile jedes einzelne Posting per ID-Referenz (vgl. Listing 3, Werte des Attributs @*who*) mit einer eindeutigen Autorenzuordnung versehen zu können. Die Trennung der Angaben zu den Interaktionsbeteiligten vom eigentlichen Logfile und die Referenzierung über IDs hat zwei Vorteile: Zum einen (1) vermeidet sie die redundante Wiederholung immer wieder derselben Beteiligteninformation bei jedem der von diesem Beteiligten produzierten Posting; zum anderen (2) ermöglicht sie bei Korpora, bei denen die Autorennamen nicht anonymisiert wurden, die Erzeugung anonymisierter Sichten, insofern sämtliche Metadaten zu den Beteiligten (hier: Name, Rolle und Geschlecht) einfach vom Logfile abgetrennt werden können; anhand der in den einzelnen Postings angegebenen Beteiligten-IDs ist in diesem Fall eine Unterscheidung der Beteiligten trotzdem noch möglich.

Im Falle des Chat-Logfiles sind die Informationen, die in der *listPerson* für die einzelnen Beteiligten gegeben werden, recht rudimentär. Sie lassen sich aber beliebig ausbauen und erweitern, beispielsweise um Angaben, die aus den Benutzerprofilen der Beteiligten extrahiert wurden (was z. B. im Falle von Online-Foren und Tweets relevant ist), oder um die Inhalte von Benutzersignaturen, die z. B. in Wikipedia-Diskussionen und in Foren den Postings der Beteiligten als automatisch aus Templates generierte Textbausteine beigestellt werden.

Die *timeline* in Listing 2 verzeichnet sämtliche für die Repräsentation des Logfiles relevante Zeitpunktangaben. Diese sind in den Ausgangsdaten den einzelnen Postings typischerweise in Form von sogenannten *Timestamps* (Zeitstempeln) zugeordnet. Die Erfassung als Teil des Headers erlaubt die Repräsentation der zeitlichen Struktur an zentralem Ort und in einem einheitlichen Format, während die Darstellung von Zeitpunktangaben in den *Timestamps* innerhalb desselben Korpus (z. B. in Chat-Logfiles aus verschiedenen Quellen) im Format variieren kann (z. B. „15. Juli 2017, 14:30 Uhr“ vs. „Sonntag, 2:30 p.m.“). Die einzelnen Postings sind den in der *timeline* beschriebenen Zeitpunktangaben (Instanzen des Elements *when*) über ID-Referenzen zugeordnet (vgl. Listing 3, Werte des Attributs @*synch*).

Listing 2: Ausschnitt 2 (gekürzt) aus dem TEI-Header für ein Dokument aus dem Dortmunder Chat-Korpus in CLARIN-D:

```
<profileDesc>
  <textClass>
    <catRef scheme="http://corpora.ids-mannheim.de/
      taxonomies/taxonomy-handlungsbereiche.tei.
      xml#handlungsbereiche"
```

```
         target="http://corpora.ids-mannheim.de/taxonomies/
         taxonomy-handlungsbereiche.tei.xml#h1204000"/>
     </textClass>
     <particDesc>
       <listPerson>
         <person role="system" xml:id="f1204009.A01_System">
           <persName type="nickname">system</persName>
           <sex evidence="estimated">system</sex>
         </person>
         <person role="expert" xml:id="f1204009.A02">
           <persName type="nickname">[_FEMALE-EXPERT-A02_]</
             persName>
           <sex evidence="estimated">female</sex>
         </person>
         <person role="client" xml:id="f1204009.A03">
           <persName type="nickname">[_MALE-CLIENT-A03_]</
             persName>
           <sex evidence="estimated">male</sex>
         </person>
       </listPerson>
     </particDesc>
    <textDesc>
       <interaction>
         <timeline>
           <when absolute="22:07:00" xml:id="f1204009.t001"/>
           <when absolute="22:08:00" xml:id="f1204009.t002"/>
           <when absolute="22:09:00" xml:id="f1204009.t003"/>
           <when absolute="22:10:00" xml:id="f1204009.t004"/>
           <when absolute="22:11:00" xml:id="f1204009.t005"/>
           <when absolute="22:12:00" xml:id="f1204009.t006"/>
           <when absolute="22:13:00" xml:id="f1204009.t007"/>
           <when absolute="22:14:00" xml:id="f1204009.t008"/>
           <when absolute="22:15:00" xml:id="f1204009.t009"/>
           <when absolute="22:16:00" xml:id="f1204009.t010"/>
         …
         </timeline>
       </interaction>
     </textDesc>
   </profileDesc>
</teiHeader>
```

Die Struktur der in den Listings 1 und 2 abgebildeten Ausschnitte aus dem TEI-Header entspricht dem TEI-Standard. Einige Elemente wurden, wie aufgezeigt, allerdings IBK-spezifisch reinterpretiert. Konkrete Änderungen des TEI-Formats gegenüber dem Standard waren auf Ebene der Strukturannotation des eigentlichen Logfiles erforderlich. Hier wurden ganz zentral Lösungen (1) für die Repräsentation von Postings und (2) für die Beschreibung von IBK-Makrostrukturen (Logfiles, Threads) benötigt. Diese Lösungen im CLARIN-D-Schema werden im Folgenden erläutert.

4.2 Annotation von Makro- und Mikrostrukturen

Die TEI-Guidelines umfassen ausgearbeitete Module für die Annotation von Strukturen in redigierten Texten und für die Annotation der Struktur transkribierter Gespräche. In Bezug auf die Frage, wie sich IBK-Strukturen sinnvoll in TEI repräsentieren lassen, ist zunächst zu klären, ob sich diese Strukturen ggf. mit in TEI bereits vorhandenen Modellen sinnvoll darstellen lassen. Drei Modellierungsoptionen aus dem TEI-Standard bieten sich grundsätzlich an und sind zu diskutieren:

– Durch die Brille *redigierter Texte* betrachtet könnten die Primärdaten von IBK-Korpora (die in Dokumenten gespeicherten Logfiles) als *Texte* betrachtet werden, die aus einer Abfolge von Einheiten bestehen, die durch Layoutmerkmale voneinander abgrenzbar sind und von verschiedenen Autoren verfasst wurden. Der Vorteil einer Entscheidung für die Beschreibung von IBK-Strukturen mit den TEI-Modellen für Textstrukturen läge darin, dass für die Beschreibung von IBK-Mikrostrukturen – z. B. die Gliederung komplexer Foren-Postings oder von Einträgen auf Wikipedia-Diskussionsseiten – in TEI ein ausgearbeitetes Inventar an Elementen bereits zur Verfügung stünde. Der Nachteil der damit einhergehenden Beschreibung von IBK-Verläufen als einer *Form von Text* wäre der, dass die Gliederung eines monologischen Textes (in Abschnitte und Absätze, TEI-Elemente `<div>` und `<p>`) typischerweise von *einer* Autor-Instanz konzipiert und verantwortet wird, während es IBK aufgrund ihres interaktionalen Charakters gerade zentral ist, dass die entstehenden Makrostrukturen sich – ähnlich wie in mündlichen Gesprächen – aus dem Zusammenwirken vieler ergeben (unähnlich zu mündlichen Gesprächen spielt, wie oben schon bemerkt, dabei aber auch die Technologie eine nicht unwesentliche Rolle).
– Durch die Brille *mündlicher Gespräche* betrachtet könnten Postings als *utterances* aufgefasst und IBK-Makrostrukturen – analog zu Gesprächs-Transkripten – als Abfolgen von *utterances* (TEI-Element `<u>`) beschrieben werden. Der Vorteil einer solchen Modellierung des Gegenstandes wäre die

Nähe des Modells zur prototypischen Form zwischenmenschlicher Interaktion. Der Nachteil bestünde zum einen darin, dass *utterances* (a) nicht als schriftliche Äußerungen konzipiert sind und (b) dass Postings sich hinsichtlich der für sie charakteristischen, konsekutiven Abfolge von Verbalisierung, Übermittlung und adressatenseitiger Verarbeitung (vgl. Abschnitt 3) fundamental anders verhalten als interaktionale mündliche Äußerungen. Das Konzept der *utterance* müsste so umfassend reinterpretiert werden, dass es für eine distinktive Erfassung von Turns in mündlicher Interaktion nicht mehr brauchbar wäre. Ebenso wie die Möglichkeit der Beschreibung von Postings als Textgliederungen (Textabschnitte oder Absätze) bedeutete eine Beschreibung von Postings als *utterances* eine unzulässige Verschleierung zentraler Charakteristika der Interaktionskonstitution in IBK.

– Durch die Brille des TEI-Modells für *performance texts*, das für die Strukturannotation von Dramentexten konzipiert ist, könnten Postings als *speeches* (TEI-Element `<sp>`) aufgefasst werden: `<sp>` beschreibt „an individual speech in a performance text, or a passage presented as such in a prose or verse text". Tatsächlich ist das Element `<sp>` das einzige TEI-Element, das für die Repräsentation interaktionaler Äußerungen vorgesehen ist, die primär (und nicht wie *utterances* in transkribierten Gesprächen erst durch sekundäre Verschriftung) medial schriftlich realisiert sind. Auch dieses Element scheidet aber aus naheliegenden Gründen für die Repräsentation von IBK-Strukturen aus: *speeches* beschreiben keine authentischen Sprachäußerungen und die durch *speeches* konstituierten Makrostrukturen sind – unter Produktionsaspekt monologisch – von einer einzigen Autor-Instanz konzipiert.

Im vorgestellten Schema wie auch schon in den beiden Vorgängerschemas haben wir uns daher dafür entschieden, für die Annotation von Postings per *customization* ein eigenes TEI-Element `<post>` einzuführen und dieses mit einem Inhaltsmodell auszustatten, das sowohl die Gemeinsamkeiten, als auch die charakteristische Differenz einerseits zu Gliederungseinheiten in monologischen Texten – also zu Einheiten, die im TEI-Universum üblicherweise als `<div>` oder `<p>` repräsentiert werden – als auch andererseits zu Sprecherbeiträgen in mündlichen Interaktionen – in TEI mit `<u>` dargestellt – zur Geltung bringt. Eventuelle Gemeinsamkeiten oder Unterschiede zu *speeches* werden nicht weiter verfolgt.

Das Element `<post>` und die durch Abfolgen von `<post>`-Instanzen konstituierten IBK-Makrostrukturen weisen die folgenden Merkmale auf:

– `<post>` ist konzipiert als ein Element vom Typ *model.divPart*. Eine Anforderung bei der *customization* besteht darin, dass neu hinzugefügte, im

Standard nicht enthaltene Elemente in das Elementklassensystem des TEI-Rahmenwerks eingeordnet werden müssen. Damit wird beschrieben, wie sich ein neu hinzugefügtes Element zu bereits vorhandenen Elementen und ihren Inhaltsmodellen verhält, welche Kind-Elemente und Attribute es von vorhandenen Modellen erbt. Durch die Konzeption von <post> als *model.divPart* wird ausgesagt, dass Postings in formaler Hinsicht Ähnlichkeiten mit Einheiten der Textgliederung aufweisen: In einem gespeicherten Logfile sind sie typischerweise als Gliederungseinheiten zu erkennen; anhand von Layouteigenschaften lassen sich Logfiles damit weitgehend automatisiert in Einheiten des Typs <post> zergliedern (vgl. die o. a. Anforderung (5)).

– Zugleich wird über das Attribut *@who* die Möglichkeit geschaffen, jedem Posting einen individuellen Autor zuzuordnen. Damit wird dargestellt, dass die IBK-Makrostruktur, die sich als Abfolge von Postings darstellt, nicht das Ergebnis einer von einer einzelnen Autor-Instanz verantworteten, monologischen Strukturbildung ist; stattdessen wird sie als interaktionale Struktur konzipiert. Dieses Merkmal teilen IBK-Makrostrukturen, vermittelt über die Konzeption des Elements <post>, mit mündlichen Gesprächen.
– Die Makrostruktur als solche wird durch das Element <div> aus dem TEI-Standard dargestellt. <div> kann Textgliederungen unterschiedlichster Art beschreiben. In den TEI-CMC-Schemas wird <div> reinterpretiert als eine Gliederungseinheit, die Sequenzen aus zwei oder mehr Instanzen des Elements <post> bündelt. Elemente des Typs <div> lassen sich über ein Attribut *@type* subklassifizieren, für das als Wertbelegungen *logfile* und *thread* empfohlen werden, um verschiedene Typen von IBK-Makrostrukturen zu unterscheiden.
– Das Element <post> lässt sich über eine Reihe von neu eingeführten oder für die Repräsentation von IBK adaptierten Attributen subklassifizieren (u. a.):
  a) Mit *@type* werden verschiedene Typen von Postings unterschieden: „Standard"-Postings in direkter Rede sowie Postings vom Typ „event", mit denen – insbesondere in Chats – eine Aussage aus „Regisseurssicht" über reale oder spielerisch vollzogene Aktivitäten eines Interaktionsbeteiligten formuliert wird („FrankieABC holt sich mal nen Kaffee", „FrankieABC betritt den Raum").
  b) *@replyTo* ermöglicht (optional) eine weitergehende Beschreibung der sequenziellen Vernetzung des Postings mit anderen Postings durch Verweis auf die entsprechenden Posting-IDs. Die Realisierung von *ReplyTo*-Beziehungen in der Annotation ist sowohl automatisiert als auch manuell denkbar: automatisiert durch automatische Auswertung

von Adressierungselementen oder Zitationen in der Mikrostruktur der Postings (vorausgesetzt, die entsprechenden Elemente sind auf der Ebene der Mikrostruktur ihrerseits annotiert); manuell als Ergebnis einer qualitativen Analyse von Postings auf dem Hintergrund ihres sequenziellen Kontexts.

c) *@auto* gibt an, ob ein Posting automatisch generiert wurde. Typische Fälle sind sogenannte „Systemmeldungen" in Chats, mit denen Resultate von Nutzereingaben, die nicht unmittelbar als Postings intendiert sind, gemeldet werden (z. B. „Friede23 betritt den Raum"). Für quantitative Korpusuntersuchungen ist es wichtig, solche Postings bei der Korpusanalyse ausfiltern zu können, da ihre Inhalte nicht auf Verbalisierungen von menschlichen Interaktionsbeteiligten zurückgehen, sondern aus von den Programmierern im System vordefinierten Templates erzeugt wurden. Das muss nicht nur auf Posting-Ebene der Fall sein; auch auf der Mikroebene von Postings können Textbausteine dieser Art vorkommen, beispielsweise in Einleitungen zu Zitatblöcken in Online-Forenbeiträgen, die über die automatische Zitierfunktion erzeugt wurden.

Listing 3 zeigt einen Ausschnitt aus einem Korpusdokument mit drei vollständig annotierten `<post>`-Instanzen, die in einem `<div>`-Element vom Typ *logfile* gebündelt sind. Die Werte zu den Attributen *@who* und *@synch* referenzieren die IDs von Einträgen in der *listPerson* und in der *timeline* aus dem TEI-Header (vgl. Listing 2 und zugehörige Erläuterungen). Auf der Mikroebene der drei abgebildeten Postings ist zudem das Ergebnis einer automatischen *Part-of-speech*-Annotation und einer Lemmatisierung, die mit den in Horbach et al. (2014) beschriebenen Sprachverarbeitungswerkzeugen erzeugt wurden, in Form einer Inline-Annotation in die TEI-Repräsentation integriert. Dazu wurde das Element `<w>` (*word*) aus dem TEI-Standard als Element für die Beschreibung morphosyntaktischer Informationen adaptiert. Die *Part-of-speech*-Informationen folgen dem für IBK-Korpora erweiterten STTS-Tagset nach Beißwenger et al. (2015).

Listing 3: Ausschnitt aus dem TEI-Body für ein Dokument aus dem Dortmunder Chat-Korpus in CLARIN-D:

```
<div type="logfile">
    <post auto="false" rend="color:black" synch="#f1204009.t040"
    type="standard" who="#f1204009.A02" xml:id="f1204009.m187">
        <time> 22:46 </time>
```

```
    <anchor type="sentence_start"/>
    <w lemma="ich" type="PPER" xml:id="f1204009.m187.
        t1">Ich</w>
    <w lemma="wünschen" type="VVFIN" xml:id="f1204009.m187.
        t2">wünsche</w>
    <w lemma="Sie|sie" type="PPER" xml:id="f1204009.m187.
        t3">Ihnen</w>
    <w lemma="alle" type="PIAT" xml:id="f1204009.m187.
        t4">alles</w>
    <w lemma="Gute" type="NN" xml:id="f1204009.m187.
        t5">Gute</w>
    <w lemma="für" type="APPR" xml:id="f1204009.m187.
        t6">für</w>
    <w lemma="ihr" type="PPOSAT" xml:id="f1204009.m187.
        t7">Ihre</w>
    <w lemma="Verhandlung" type="NN" xml:id="f1204009.m187.
        t8">Verhandlungen</w>
    <w lemma="mit" type="APPR" xml:id="f1204009.m187.
        t9">mit</w>
    <w lemma="die" type="ART" xml:id="f1204009.m187.
        t10">dem</w>
    <w lemma="DPD" type="NN" xml:id="f1204009.m187.
        t11">DPD</w>
</post>
<post auto="false" rend="color:black" synch="#f1204009.t040"
type="standard" who="#f1204009.A02" xml:id="f1204009.m188">
    <time> 22:46 </time>
    <anchor type="sentence_start"/>
    <w lemma="und" type="KON" xml:id="f1204009.m188.
        t1">und</w>
    <w lemma="wünschen" type="VVFIN" xml:id="f1204009.m188.
        t2">wünsche</w>
    <w lemma="Sie|sie" type="PPER" xml:id="f1204009.m188.
        t3">Ihnen</w>
    <w lemma="alle" type="PIAT" xml:id="f1204009.m188.
        t4">alles</w>
    <w lemma="Gute" type="NN" xml:id="f1204009.m188.
        t5">Gute</w>
    <w lemma="für" type="APPR" xml:id="f1204009.m188.
        t6">für</w>
```

```
        <w lemma="die" type="ART" xml:id="f1204009.m188.
            t7">den</w>
        <w lemma="zukünftig" type="ADJA" xml:id="f1204009.m188.
            t8">zukünftigen</w>
        <w type="NN" xml:id="f1204009.m188.
            t9">Online-Handel</w>
        <w lemma="!" type="$." xml:id="f1204009.m188.
            t10">!</w>
    </post>
    <post auto="false" rend="color:black" synch="#f1204009.t040"
    type="standard" who="#f1204009.A03" xml:id="f1204009.m189">
        <time> 22:46 </time>
        <anchor type="sentence_start"/>
        <w lemma="ja" type="PTKANT" xml:id="f1204009.m189.
            t1">Ja</w>
        <w lemma="," type="$," xml:id="f1204009.m189.
            t2">,</w>
        <w lemma="die" type="PDS" xml:id="f1204009.m189.
            t3">das</w>
        <w lemma="können" type="VMFIN" xml:id="f1204009.m189.
            t4">kann</w>
        <w lemma="ich" type="PPER" xml:id="f1204009.m189.
            t5">ich</w>
        <w lemma="gebrauchen" type="VVINF" xml:id="f1204009.
            m189.t6">gebrauchen</w>
        <w type="EMOASC" xml:id="f1204009.m189.t7">:-(</w>
    </post>
…
</div>
```

Im Projekt *ChatCorpus2CLARIN* wurde das vorgestellte Schema für die Remodellierung des kompletten Dortmunder Chat-Korpus (1 Million Tokens) in TEI angewendet (Lüngen et al. 2016). Neben Chat-Daten flossen in die Entwicklung des Schemas Analysen und Annotationsexperimente zu einem Datenset mit Stichproben weiterer IBK-Formen ein (Wikipedia-Diskussionsseiten, WhatsApp-Interaktionen, Newskommunikation, Tweets). Die Vorgängerversion von Beißwenger et al. (2012) wurde von Margaretha & Lüngen (2014) für die Repräsentation der Posting-Struktur im Wikipedia-Diskussionsseiten-Korpus in DeReKo adaptiert. Die Vorgängerversion von Chanier et al. (2014) lag der Strukturannotation von vierzehn französischen IBK-Korpora zu neun unter-

schiedlichen IBK-Formen zugrunde (u. a. SMS, Wikipedia-Diskussionen, Tweets, Weblogs, E-Mails, Foren, Chats). Die aktuelle Schemaversion wird gegenwärtig an der Universität Gießen für die Strukturannotation eines Scienceblog-Korpus eingesetzt (Grumt Suárez, Karlova-Bourbonus & Lobin 2016). In den genannten Projekten hat sich das Basisschema als praktikabel erwiesen. Im Rahmen von *ChatCorpus2CLARIN* hat die Verwendung des Schemas eine Integration der Ressource in vorhandene Korpussammlungen zur geschriebenen deutschen Sprache ermöglicht, die bereits in Standard-TEI repräsentiert sind (DeReKo, DWDS).

5 Fazit und Ausblick

Die Sprachressourcen-Infrastruktur der Zukunft wird IBK-Korpora umfassen, die
- eine breite Zahl von IBK-Genres und Sprachen abdecken,
- frei für Forschung und Lehre zur Verfügung stehen, getreu dem Motto der europäischen Sprachressourcen-Infrastruktur-Initiative CLARIN „Sprachressourcen für alle“ (*Common Language Resources*),
- für linguistische Analysezwecke aufbereitet, d. h. um linguistische Strukturannotationen angereichert und durch Metadaten erschlossen sind,
- in Übereinstimmung mit Standards im Bereich der Digital Humanities repräsentiert sind,
- interoperabel sowohl untereinander als auch mit Korpora anderen Typs (Textkorpora, Korpora gesprochener Sprache) sind und sich daher mit denselben Korpusrecherchewerkzeugen vergleichend mit anderen Korpora auswerten lassen.

Diese Vision ist nicht unrealistisch: In einer wachsenden Zahl von Projekten zu unterschiedlichen Sprachen entstehen derzeit Korpora zur internetbasierten Kommunikation und werden Lösungen für die mit diesem Korpustyp verbundenen Desiderate entwickelt. Standardisierungs- und Infrastrukturinitiativen im Bereich der Digital Humanities haben den Bedarf an Forschungsressourcen zur internetbasierten Kommunikation erkannt und unterstützen die Entwicklung von Lösungen, um sie in die Ressourcenlandschaft einzugliedern und mit existierenden Ressourcen zu vernetzen.

Ein wichtiger Baustein beim Aufbau interoperabler IBK-Korpora ist ein Standard für die Repräsentation von IBK-Daten. Die TEI bietet ein flexibles Rahmenwerk, um einen solchen Standard zu entwickeln. Die von der TEI-CMC-SIG vorgelegten Entwürfe stehen als vollständige TEI-Schemas zur Verfügung, wurden begleitend zu ihrer Entwicklung in verschiedenen Korpusprojekten er-

probt und können von jedermann für die Annotation eigener Korpora genutzt werden. Nach wie vor handelt es sich bei diesen Schemas um *customizations*: Sie sind mit dem TEI-Standard vollumfänglich kompatibel, enthalten aber dennoch einzelne Modelle, die selbst nicht Teil des Standards sind. Ein wichtiger nächster Schritt in der Arbeit der CMC-SIG, in der Akteure von Korpusprojekten zu unterschiedlichen Sprachen mitwirken, besteht daher darin, aus den bislang vorgelegten *customizations* und den mit ihrer Anwendung gemachten Erfahrungen Eingaben für den TEI-Standardisierungsprozess zu formulieren.

Ein Standard für die Repräsentation von IBK muss sich notwendigerweise auf die Bereitstellung von Modellen beschränken, die basale Interaktionsformate darstellbar machen, die in vielen Kommunikationsumgebungen im Internet eine Rolle spielen und die sich als relativ stabil gegenüber dem beständigen technologischen Wandel erweisen. Das in diesem Beitrag beschriebene Posting-basierte Interaktionsformat, das zahlreichen Formen internetbasierter Kommunikation zugrundeliegt, ist ein Beispiel für ein solches basales Format. In einem weiteren Schritt wurde gezeigt, wie sich dieses Format mit dem aktuellen Schemaentwurf der TEI-CMC-SIG in einer XML-Repräsentation darstellen lässt, die mit dem TEI-Standard kompatibel ist. Eine künftige Aufgabe der CMC-SIG wird es sein, dieses Basisformat, das vor allem auf schriftbasierte Formen internetbasierter Kommunikation fokussiert, um Modelle für die Repräsentation von Daten aus multimodalen IBK-Umgebungen zu erweitern. Erste Ansätze dafür, wenn auch im vorliegenden Beitrag nicht näher behandelt, sind im CoMeRe-Schema (Chanier et al. 2014) angelegt und wurden im CLARIN-D-Schema fortgeschrieben. Ein drittes wichtiges Desiderat stellt die Anpassung von Schemata für die Erfassung von Metadaten für den Bereich der IBK dar: Gerade weil sich Kommunikationstechnologien und -umgebungen beständig wandeln, müssen Beschreibungen zur Struktur und zum Funktionsumfang von Kommunikationsumgebungen, aus denen Daten erhoben worden, in einer Weise erfasst und repräsentiert werden, dass Nutzer der entsprechenden Korpora in 10 oder 20 Jahren, möglicherweise auch schon in 3 Jahren, noch rekonstruieren können, welche technologischen Rahmenbedingungen die sprachliche Gestaltung von Interaktion, so wie es sich in den auch Daten zeigt, beeinflusst haben.

Ein Standard für die Repräsentation und Strukturannotation stellt dabei letzten Endes nur *einen* Baustein für die Verbesserung der Ressourcenlage in Bezug auf die Domäne der internetbasierten Kommunikation dar. Um die als Vision skizzierte IBK-Korpuslandschaft der Zukunft zu realisieren, müssen Innovationen bei der sprachtechnologischen Verarbeitung von IBK-Daten und bei der Anpassung von Werkzeugen für die Korpusrecherche und -analyse, *best practices* für die Adressierung juristischer und forschungsethischer Fragen bei der Datenerhebung und -dokumentation sowie Lösungen für die Anonymisierung der Korpusdaten als weitere Bausteine hinzutreten.

Die Korpora internetbasierter Kommunikation von morgen bilden einen (möglicherweise bedeutsamen) Teil der kulturellen Überlieferung des Alltagssprachgebrauchs von heute für die Sprachhistoriker von übermorgen.[33] Die daraus resultierenden Anforderungen an die Bereitstellung und Repräsentation von Korpora ist damit nicht nur gegenwartsbezogen, sprachen- und domänenübergreifend, sondern auch diachron bezogen eine wichtige Aufgabe bei der Arbeit an der Sprachressourceninfrastruktur der Zukunft.

Literatur

Auer, Peter (2000): On line-Syntax – oder: was es bedeuten könnte, die Zeitlichkeit der mündlichen Sprache ernst zu nehmen. In *Sprache und Literatur* 85, 43–56.

Barbaresi, Adrien (2016): Efficient construction of metadata-enhanced web corpora. In *Proceedings of the 10th Web as Corpus Workshop, Association for Computational Linguistics*, 7–16. https://hal.archives-ouvertes.fr/hal-01371704v2/document (letzter Zugriff: 8. 11. 2017).

Barbaresi, Adrien & Kay-Michael Würzner (2014): For a fistful of blogs: Discovery and comparative benchmarking of republishable German content. In *Proceedings of NLP4CMC workshop (KONVENS 2014)*, 2–10. Hildesheim University Press. https://www.dwds.de/static/publications/pdf/Barbaresi-Wuerzner_Fistful-of-blogs_2014.pdf (letzter Zugriff: 8. 11. 2017).

Beißwenger, Michael (2003): Sprachhandlungskoordination im Chat. *Zeitschrift für germanistische Linguistik* 31 (2), 198–231.

Beißwenger, Michael (2007): *Sprachhandlungskoordination in der Chat-Kommunikation*. Berlin, New York: de Gruyter (Reihe Linguistik – Impulse & Tendenzen 26).

Beißwenger, Michael (2010): Chattern unter die Finger geschaut: Formulieren und Revidieren bei der schriftlichen Verbalisierung in synchroner internetbasierter Kommunikation. In Vilmos Ágel & Mathilde Hennig (Hrsg.), *Nähe und Distanz im Kontext variationslinguistischer Forschung*, 247–294. Berlin, New York: de Gruyter.

Beißwenger, Michael (2013): Das Dortmunder Chat-Korpus. In *Zeitschrift für germanistische Linguistik* 41 (1), 161–164.

Beißwenger, Michael (2016): Praktiken in der internetbasierten Kommunikation. In Arnulf Deppermann, Helmuth Feilke & Angelika Linke (Hrsg.), *Sprachliche und kommunikative Praktiken*. Jahrbuch 2015 des Instituts für Deutsche Sprache, 279–310. Berlin/New York: de Gruyter.

Beißwenger, Michael (Hrsg.) (2017): *Empirische Erforschung internetbasierter Kommunikation*. Berlin, New York: de Gruyter (Empirische Linguistik/Empirical Linguistics 9).

33 Dieser Gedanke geht auf einen Diskussionsbeitrag von Erhard Hinrichs im Rahmen eines deutsch-französischen Kolloquiums zu Standards für IBK-Korpora an der Universität Duisburg-Essen zurück (Gedächtniszitat). Ich danke Erhard Hinrichs für diese weitsichtige Anregung.

Beißwenger, Michael & Angelika Storrer (2008): Corpora of computer-mediated communication. In Anke Lüdeling & Merja Kytö (Hrsg.), *Corpus Linguistics HSK* 29 (1), 292–309. Berlin: de Gruyter.

Beißwenger, Michael & Angelika Storrer (2012): Interaktionsorientiertes Schreiben und interaktive Lesespiele in der Chat-Kommunikation. *Zeitschrift für Literaturwissenschaft und Linguistik* 168, 92–124.

Beißwenger, Michael & Lothar Lemnitzer (2013): Aufbau eines Referenzkorpus zur deutschsprachigen internetbasierten Kommunikation als Zusatzkomponente für die Korpora im Projekt „Digitales Wörterbuch der deutschen Sprache" (DWDS). *Journal for Language Technology and Computational Linguistics* 28 (2), 1–22.

Beißwenger, Michael, Nelleke Oostdijk, Angelika Storrer & Henk van den Heuvel (2014): Building and Annotating Corpora of Computer-Mediated Communication: Issues and Challenges at the Interface of Corpus and Computational Linguistics. Journal of Language Technology and Computational Linguistics 2. http://jlcl.org (letzter Zugriff: 8. 11. 2017).

Beißwenger, Michael, Thomas Bartz, Angelika Storrer & Swantje Westpfahl (2015): Tagset und Richtlinie für das Part-of-Speech-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation. Guideline-Dokument aus dem Projekt „GSCL Shared Task: Automatic Linguistic Annotation of Computer-Mediated Communication/Social Media" (EmpiriST 2015). http://ttps//sites.google.com/site/empirist2015/home/annotation-guidelines (letzter Zugriff: 8. 11. 2017).

Beißwenger, Michael, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer & Angelika Storrer (2012): A TEI Schema for the Representation of Computer-mediated Communication. *Journal* of *the Text Encoding Initiative* 3. http://jtei.revues.org/476 (doi:10.4000/jtei.476) (letzter Zugriff: 8. 11. 2017).

Beißwenger, Michael, Thierry Chanier, Tomaž Erjavec, Darja Fišer, Axel Herold, Nikola Lubešic, Harald Lüngen, Céline Poudat, Egon Stemle, Angelika Storrer & Ciara Wigham (2017a): Closing a Gap in the Language Resources Landscape: Groundwork and Best Practices from Projects on Computer-mediated Communication in four European Countries. *Selected Papers from the CLARIN Annual Conference 2016*, October 26–28 2016, France, Aix-en-Provence. Linköping: Linköping University Electronic Press (Linköping University Electronic Conference Proceedings), 1–18.

Beißwenger, Michael, Harald Lüngen, Jan Schallaböck, John H. Weitzmann, Axel Herold, Pawel Kamocki, Angelika Storrer & Julia Wildgans (2017b): Rechtliche Bedingungen für die Bereitstellung eines Chat-Korpus in CLARIN-D: Ergebnisse eines Rechtsgutachtens. In Michael Beißwenger (Hrsg.), *Empirische Erforschung internetbasierter Kommunikation*. Berlin, New York: de Gruyter (Empirische Linguistik/Empirical Linguistics 9), 7–46.

Bolander, Brook & Miriam A. Locher (2014): Doing Sociolinguistic Research on Computer-Mediated Data: A Review of Four Methodological Issues. *Discourse, Context & Media* (3), 14–26.

Brinker, Klaus, Hermann Cölfen & Steffen Pappert (2014): *Linguistische Textanalyse: eine Einführung in Grundbegriffe und Methoden*. 8., neu bearb. und erw. Aufl. Berlin: Erich Schmidt.

Chanier, Thierry, Celine Poudat, Benoit Sagot, Georges Antoniadis, Ciara Wigham, Linda Hriba, Julien Longhi & Djamé Seddah (2014): The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. Journal of language Technology and

Computational Linguistics 29 (2), 1–30. http://www.jlcl.org/2014_Heft2/1Chanier-et-al.pdf (letzter Zugriff: 8. 11. 2017).
Cherny, Lynn (1999): *Conversation and Community. Chat in a Virtual World.* Stanford: University of Chicago Press.
Chiari, Isabella & Alessio Canzonetti (2014): Le forme della comunicazione mediata dal computer: generi, tipi e standard di annotazione. In Enrico Garavelli & Elina Suomela-Härmä (Hrsg.), *Dal manoscritto al web: canali e modalità di transmissione dell'italiano. Tecniche, materiali e usi nella storia della lingua*. Atti del XII Convegno della Società Internazionale di Linguistica e Filologia Italiana (SILFI), 18–19 June 2012, Helsinki, 595–606. Florenz: Franco Cesati Editore.
CLARIN-D AP 5 (2012): *CLARIN-D User Guide*. Version: 1.0.1. https://www.clarin-d.de/en/help/user-handbook (letzter Zugriff: 8. 11. 2017).
Deppermann, Arnulf, Helmuth Feilke & Angelika Linke (Hrsg.) (2016): *Sprachliche und kommunikative Praktiken*. Jahrbuch 2015 des Instituts für Deutsche Sprache. Berlin/New York: de Gruyter.
DiDi (2015): Beschreibung der Anonymisierung im DiDi-Korpus. http://www.eurac.edu/en/research/autonomies/commul/Documents/DiDi/DiDi_anonymisation_DE.pdf (letzter Zugriff: 8. 11. 2017).
Dürscheid, Christa (2005): Medien, Kommunikationsformen, kommunikative Gattungen. *Linguistik online* 22 (1). http://www.linguistik-online.de/22_05/duerscheid.pdf (letzter Zugriff: 8. 11. 2017).
Dürscheid, Christa & Elisabeth Stark (2011): sms4science: An international corpus-based texting project and the specific challenges for multilingual Switzerland. In Crispin Thurlow & Kristine Mroczek (Hrsg.), *Digital Discourse. Language in the New Media*, 299–320. Oxford: Oxford University Press.
Ehlich, Konrad (1984): Zum Textbegriff. In Annely Rothkegel & Barbara Sandig (Hrsg.), *Text – Textsorten – Semantik*, 531–550. Hamburg: Buske.
Fišer, Darja, Tomaž Erjavec & Nikola Ljubešić (2016): JANES v0.4: Korpus slovenskih spletnih uporabniških vsebin. *Slovenščina* 2.0 4(2), 67–99. http://dx.doi.org/10.4312/slo2.0.2016.2.67-99 (letzter Zugriff: 8. 11. 2017).
Fišer, Darja & Michael Beißwenger (Hrsg.) (2017): *Investigating Computer-Mediated Communication: Corpus-Based Approaches to Language in the Digital World*. Ljubljana: Ljubljana University Press (Translation Studies and Applied Linguistics). Open Access: https://knjigarna.ff.uni-lj.si/en/izdelek/1766/investigating-computer-mediated-communication/
Fišer, Darja, Tomaž Erjavec & Nikola Ljubešić (2017): The compilation, processing and analysis of the Janes corpus of Slovene user-generated content: In Ciara R Wigham & Gudrun Ledegen (Hrsg.), *Corpus de Communication Médiée par les Réseaux. Construction, structuration, analyse*, 125–138. Paris: L'Harmattan (Humanités numériques).
Forsyth, Eric N. & Craig H. Martell (2007): Lexical and Discourse Analysis of Online Chat Dialog. In *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007)*, USA, Irvine, 19–26, https://catalog.ldc.upenn.edu/docs/LDC2010T05/lex_anal_online_chat.pdf (letzter Zugriff: 8. 11. 2017).
Frey, Jennifer-Carmen, Egon W. Stemle & Aivars Glaznieks (2014): Collecting Language Data of Non-Public Social Media Profiles. In Gertrud Faaß & Josef Ruppenfofer (Hrsg.), *Workshop Proceedings of the 12th Edition of the KONVENS Conferenc*e, 11–15. Hildesheim: Universitätsverlag Hildesheim.

Frey, Jennifer-Carmen, Aivars Glaznieks & Egon W. Stemle (2016): The DiDi Corpus of South Tyrolean CMC Data: A Multilingual Corpus of Facebook Texts. In *Proceedings of the Third Italian Conference on Computational Linguistics* (CLiC-it 2016), 5–6 December 2016, Italy, Napoli, http://ceur-ws.org/Vol-1749/paper27.pdf (letzter Zugriff: 8. 11. 2017).

Garcia, Angela Cora & Jennifer Baker Jacobs (1998): The Interactional Organization of Computer Mediated Communication in the College Classroom. *Qualitative Sociology* 21 (3), 299–317.

Garcia, Angela Cora & Jennifer Baker Jacobs (1999): The Eyes of the Beholder: Understanding the Turn-Taking System in Qua-si-Synchronous Computer-Mediated Communication. *Research on Language and Social Interaction* 32 (4), 337–367.

Geyken, Alexander, Adrien Barbaresi, Jörg Didakowski, Bryan Jurish, Frank Wiegand & Lothar Lemnitzer (2017): Die Korpusplattform des „Digitalen Wörterbuchs der deutschen Sprache" (DWDS). *Zeitschrift für germanistische Linguistik* 45 (2), 327–344.

Giesbrecht, Eugenie & Stefan Evert (2009): Is Part-of-Speech Tagging a Solved Task? An Evaluation of POS Taggers for the German Web as Corpus. In *Proceedings of the 5th Web as Corpus Workshop (WAC5)*. San Sebastian, Spain. http://www.stefan-evert.de/PUB/GiesbrechtEvert2009_Tagging.pdf (letzter Zugriff: 8. 11. 2017).

Grumt Suárez, Holger, Natali Karlova-Bourbonus & Henning Lobin (2016): Compilation and Annotation of the Discourse-structured Blog Corpus for German. *In Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanitie*s (cmc-corpora2016), University of Ljubljana. http://nl.ijs.si/janes/wp-content/uploads/2016/09/CMC-2016_Grumt_et_al_Compilation-and-Annotation.pdf (letzter Zugriff: 8. 11. 2017).

Herring, Susan C. (1996): *Computer-Mediated Communication. Linguistic, Social and Cross-Cultural Perspectives*. Amsterdam, Philadelphia: John Benjamins Publishing Company (Pragmatics & Beyond New Series 39).

Herring, Susan C. (1999): Interactional Coherence in CMC. *Journal of Computer-Mediated Communication* 4 (4), doi:10.1111/j.1083–6101.1999.tb00106.x (letzter Zugriff: 8. 11. 2017).

Horbach, Andrea, Diana Steffen, Stefan Thater & Manfred Pinkal (2014): Improving the Performance of Standard Part-of-Speech Taggers for Computer-Mediated Communication. In *Proceedings of KONVENS 2014*, 171–177. Hildesheim: Universitätsverlag Hildesheim.

Horsmann, Tobias & Torsten Zesch (2015): Effectiveness of Domain Adaptation Approaches for Social Media PoS Tagging. In *Proceeding of the Second Italian Conference on Computational Linguistics*, 166–170. Trento: Accademia University Press.

Imo, Wolfgang (2015a): Vom ikonischen über einen indexikalischen zu einem symbolischen Ausdruck? Eine konstruktionsgrammatische Analyse des Emoticons :-) In Kerstin Fischer, Anatol Stefanowitsch, Alexander Lasch & Jörg Bücker (Hrsg.), *Konstruktionsgrammatik 5. Konstruktionen im Spannungsfeld von sequenziellen Mustern, kommunikativen Gattungen und Textsorten*, 133–162. Tübingen: Stauffenburg-Verlag.

Imo, Wolfgang (2015b): Vom Happen zum Häppchen … Die Präferenz für inkrementelle Äußerungsproduktion in internetbasierten Messengerdiensten. *Networx* 69, http://www.mediensprache.net/de/networx/networx-69.aspx (letzter Zugriff: 8. 11. 2017).

iRights.Law Rechtsanwälte (2016): *Rechtsgutachten zur Integration mehrerer Text-Korpora in die CLARIN-D-Infrastrukturen*. (Manuskript, 46 Seiten).

JIM-Studie (2016): Jugend, Information, (Multi-)Media. Basisuntersuchung zum Medienumgang 12–19-Jähriger. Hrsg. v. Medienpädagogischen Forschungsverbund Südwest. http://www.mpfs.de/de/studien/jim-studien/2016/

Jucker, Andreas H. & Christa Dürscheid (2012). The Linguistics of Keyboard-to-screen Communication. A New Terminological Framework. *Linguistik Online* 56, 39–64.
König, Katharina (2015): Dialogkonstitution und Sequenzmuster in der SMS- und WhatsApp-Kommunikation. *Travaux neuchâtelois de linguistique* 63, 87–107.
Lemnitzer, Lothar & Heike Zinsmeister (2015): *Korpuslinguistik. Eine Einführung*. 3., überarb. u. rew. Aufl. Tübingen: Narr (Narr Studienbücher).
Lindemann, Katrin, Emanuel Ruoss & Caroline Weinzinger (2014): Dialogizität und sequenzielle Verdichtung in der Forenkommunikation: Editieren als kommunikatives Verfahren. *Zeitschrift für Germanistische Linguistik* 42 (2), 223–252.
Ljubešić, Nikola, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak & Iza Škrjanec (2015): Predicting the level of text standardness in user-generated content. In *Proceedings of the 10th International Conference on Recent Advances in Natural Language Processing*, Bulgaria, Hissar, 371–378.
Lobin, Henning (2010): *Computerlinguistik und Texttechnologie*. München: Fink.
Lüngen, Harald & C. M. Sperberg-McQueen (2012): A TEI P5 Document Grammar for the IDS Text Model. *Journal of the Text Encoding Initiative (jTEI)* 3, http://jtei.revues.org/508 (letzter Zugriff: 8. 11. 2017).
Lüngen, Harald, Michael Beißwenger, Axel Herold & Angelika Storrer (2016): Integrating corpora of computer-mediated communication in CLARIN-D: Results from the curation project ChatCorpus2CLARIN. In Stefanie Dipper, Friedrich Neubarth & Heike Zinsmeister (Hrsg.), *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, 1561–64.
Lüngen, Harald (2017): DeReKo – Das Deutsche Referenzkorpus. Schriftkorpora der deutschen Gegenwartssprache am Institut für Deutsche Sprache in Mannheim. *Zeitschrift für germanistische Linguistik* 45 (1), 161–170.
Margaretha, Eliza & Harald Lüngen (2014): Building Linguistic Corpora from Wikipedia Articles and Discussions. *Journal of language Technology and Computational Linguistics* 29 (2), 59–82. http://www.jlcl.org/2014_Heft2/3MargarethaLuengen.pdf (letzter Zugriff: 8. 11. 2017).
Markman, Kris (2006): *Computer-Mediated Conversation: The Organization of Talk in Chat-Based Virtual Team Meetings*. Dissertation. University Texas at Austin.
Marx, Konstanze (2015): „kümmert euch doch um euren Dreck“ – Verteidigungsstrategien im Cybermobbing dargestellt an einem Beispiel der Plattform Isharegossip.com. In U. Tuomarla et al. (Hrsg.), *Misskommunikation und Gewalt. Mémoires de la Société Néophilologique de Helsinki*, 125–138. Vantaa: Hansaprint Oy.
Murray, Denise E. (1989): When the medium determines turns: turn-taking in computer conversation. In Hywel Coleman (Hrsg.), *Working with Language. A Multidisciplinary consideration of Language Use in Work Contexts*, 319–337. Berlin, New York: de Gruyter Mouton.
Oostdijk, Nelleke, Martin Reynaert, Véronique Hoste & Ineke Schuurman (2013): The Construction of a 500 Million Word Reference Corpus of Contemporary Written Dutch. In Peter Spyns & Jan Odijk (Hrsg.), *Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme*, 219–247. Berlin: Springer Verlag.
Perkuhn, Reiner, Holger Keibel & Marc Kupietz (2012): *Korpuslinguistik*. München: Fink.
Schönfeldt, Juliane & Andrea Golato (2003): Repair in Chats: A Conversation Analytic Approach. *Research on Language and Social Interaction* 36 (3), 241–284.
Schröck, Jasmin & Harald Lüngen (2015): Building and Annotating a Corpus of German-Language Newsgroups. In *Proceedings of the 2nd Workshop on Natural Language*

Processing for Computer-Mediated Communication/Social Media (NLP4CMC2015). Germany, Essen, 17–22. https://sites.google.com/site/nlp4 cmc2015/program (letzter Zugriff: 8. 11. 2017).

Selting, Margaret & Elizabeth Couper-Kuhlen (2000): Argumente für die Entwicklung einer interaktionalen Linguistik. *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 1, 76–95. http://www.gespraechsforschung-ozs.de (letzter Zugriff: 8. 11. 2017).

Severinson Eklundh, Kerstin (2010): To Quote or Not to Quote: Setting the Context for Computer-Mediated Dialogues. *Language@Internet* 7. http://www.languageatinternet.org/articles/2010/2665 (letzter Zugriff: 8. 11. 2017).

Stertkamp, Wolf (2016): *Spiel, Satz, Sieg: Sprache und Kommunikation in Online-Computerspielen. Eine qualitative Analyse multimodaler Kommunikation in Massively Multiplayer Online Role-Playing Games am Beispiel von Word of Warcraft*. Dissertation. Justus-Liebig-Universität Gießen.

Storrer, Angelika (2001): Sprachliche Besonderheiten getippter Gespräche: Sprecherwechsel und sprachliches Zeigen in der Chat-Kommunikation. In Michael Beißwenger (Hrsg.), *Chat-Kommunikation. Sprache, Interaktion, Sozialität und Identität in synchroner computervermittelter Kommunikation. Perspektiven auf ein interdisziplinäres Forschungsfeld*, 3–24. Stuttgart: ibidem.

Storrer, Angelika (2011): Korpusgestützte Sprachanalyse in Lexikographie und Phraseologie. In Karlfried Knapp (Hrsg.), *Angewandte Linguistik*. 3., vollst. überarb. und erw. Aufl., 216–239. Tübingen: Francke.

Storrer, Angelika (2014): Sprachverfall durch internetbasierte Kommunikation? Linguistische Erklärungsansätze – empirische Befunde. In Albrecht Plewina & Andreas Witt (Hrsg.), *Sprachverfall? Dynamik – Wandel – Variation. Jahrbuch des Instituts für Deutsche Sprache 2013*, 171–196. Berlin, Boston: de Gruyter.

Storrer, Angelika (2018): Interaktionsorientiertes Schreiben im Internet. In: Deppermann, Arnulf (Hg.): Sprache im kommunikativen, interaktiven und kulturellen Kontext, 219–244. Berlin, New York: de Gruyter.

TEI Consortium (2007): *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. http://www.tei-c.org/Guidelines/P5/ (letzter Zugriff: 8. 11. 2017).

Verheijen, Lieke & Wessel Stoop (2016): Collecting Facebook Posts and WhatsApp Chats. In *Proceedings. Text, Speech, and Dialogue: 19th International Conference*, Czech Republic, Brno, September 12–16, 249–58. Cham: Springer International Publishing.

[WAC-X/EmpiriST 2016] *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*. Stroudsburg: Association for Computational Linguistics (ACL Anthology W16–26). http://aclweb.org/anthology/W/W16/W16-26.pdf (letzter Zugriff: 8. 11. 2017).

Wigham, Ciara R. & Gudrun Ledegen (Hrsg.) (2017): *Corpus de Communication Médiée par les Réseaux*. Paris: L'Harmattan.

Gerhard Heyer, Gregor Wiedemann und Andreas Niekler

15 Topic-Modelle und ihr Potenzial für die philologische Forschung

Abstract: Statistische Analyseverfahren, die über das bloße Auszählen von Sprachdaten hinausgehen, gewinnen in der Korpuslinguistik zunehmend an Bedeutung. Der Beitrag behandelt das sogenannte *Topic Modelling* als eine besonders prominente Klasse statistischer Verfahren. Im Mittelpunkt steht der grundlegende Modellierungsansatz von Topic-Modellen, der am Beispiel der *Latent Dirichlet Allocation* (LDA) präsentiert wird. Beispielanwendungen auf verschiedenen Textquellen werden vorgestellt und dabei auch die Aspekte Reproduzierbarkeit und Reliabilität der Ergebnisse diskutiert.

Keywords: Architekturen, Evolutionäre Algorithmen, Forschungswerkzeuge, Texttechnologie

1 Einleitung

Im fortgeschrittenen Zeitalter der Digitalisierung sind über die letzten Jahrzehnte die uns zur Verfügung stehenden digitalen Textquellen in Größenordnungen gewachsen, die sich nur noch schwer fassen lassen. Dies betrifft sowohl die Retro-Digitalisierung von zunächst als Druckversion erschienenen Texten, als auch neue, primär oder ausschließlich für das digitale Medium erstellte Texte, sogenannte *born digital documents*, wie etwa Microblogs oder Texte in sozialen Netzwerken. Aufsehen erregte 2011 das Projekt *Culturomics*, für das von Geisteswissenschaftlern in Zusammenarbeit mit Google ein digitales Korpus aus Millionen von Büchern, ca. 4 % aller jemals gedruckten Titel, erstellt wurde (Michel et al. 2011). Auch rein deutsche digitale Textsammlungen wie das *Deutsche Textarchiv* mit ca. 145 Millionen beziehungsweise das

Gerhard Heyer, Universität Leipzig, Abteilung Automatische Sprachverarbeitung, Augustusplatz 10, D-04109 Leipzig, E-Mail: heyer@informatik.uni-leipzig.de
Gregor Wiedemann, Universität Leipzig, Abteilung Automatische Sprachverarbeitung, Augustusplatz 10, D-04109 Leipzig, E-Mail: gregor.wiedemann@informatik.uni-leipzig.de
Andreas Niekler, Universität Leipzig, Abteilung Automatische Sprachverarbeitung, Augustusplatz 10, D-04109 Leipzig, E-Mail: aniekler@informatik.uni-leipzig.de

https://doi.org/10.1515/9783110538663-016

Deutsche Referenzkorpus mit ca. 31 Milliarden laufenden Wortformen haben mittlerweile einen beachtlichen Umfang erreicht. Die zeitgenössische Quellenlage dieses „digitalen Überflusses" eröffnet den Sprach- und Literaturwissenschaften nicht nur neue Möglichkeiten empirischen Forschens, sie zwingt sie gerade dazu, wenn die mühsam aufgebauten digitalen Äquivalente des Archivs nicht in ähnlicher Weise verstauben sollen, wie ihre analogen Vorgänger. Diese Problematik aufgreifend fragte der Altphilologe Gregory Crane (2006): „What do you do with a million books?" Neben der digitalen Aufbereitung für Suche und Darstellung einzelner Werke liegt die digitale, computergestützte Auswertung ganzer Kollektionen als Antwort nahe. 2007 formulierte der Italiener Franco Moretti in diesem Sinne sein Forschungsprogramm des *Distant Reading* aus (2007). Anstelle einen kleinen Korpus nahezu willkürlich selektierter Klassiker immer und immer wieder durch intensives Lesen zu studieren, sollten sich Literaturwissenschaftler der „Weltliteratur" annehmen, in dem sie viele tausend Werke gleichzeitig und aus gewissem Abstand vergleichend untersuchen. Für Moretti war klar, dass dies nur mit (text-)statistischen Verfahren und computergestützten Visualisierungen gelingen kann.

Seit einigen Jahren nun, in denen die Digital Humanities erblühen, führt die digitale Quellenlage Literatur- und Sprachwissenschaftler, (Computer-)Linguisten und Informatiker enger zusammen in ihren Bemühungen, den sehr großen Textmengen Herr zu werden. Grundlage dafür sind neben den in der Syntaxanalyse traditionell weit verbreiteten musterbasierten Ansätzen insbesondere statistische Modelle und Verfahren für die automatische und semiautomatische Sprachanalyse, welche eine (qualitative) Auswertung sehr großer Textmengen auf quantitativer Basis ermöglichen. Eine besonders prominente Klasse statistischer Verfahren, die im Folgenden genauer ausgeführt werden soll, ist das sogenannte *Topic Modelling*, mit dem eine große Dokumentmenge vollkommen automatisch in unterschiedliche Themenbereiche unterteilt werden kann (Blei 2012). Die Modellierungsannahme ist, dass jede Wortform einem Themenbereich zugehörig ist und sich aus der Verteilung thematisch zusammengehöriger Wortformen in einem Text die zugrundeliegenden Themenbereiche, eben die *Topics*, ableiten lassen. Große Textkollektionen können damit nicht nur thematisch unterteilt werden, sondern Texte beziehungsweise Textabschnitte können auch anhand der identifizierten Themenstruktur klassifiziert und zusammengefasst werden. Topic-Modelle ermöglichen somit eine primär inhaltsgeleitete Strukturierung von Texten. Auch wenn sie nicht direkt zur Extraktion individueller (Syntax-)Strukturen, Informationen oder Aussageregelmäßigkeiten beitragen, können die Ergebnisse der automatischen thematischen Einteilung als Vorverarbeitungsschritt für die weitergehende Analyse sehr nützlich sein. Insofern Topic-Modelle inhaltlich interpretierbare Cluster

rein aus statistischer Beobachtung von Regelmäßigkeiten in sprachlichen Oberflächenstrukturen ableiten, nämlich dem gemeinsamen Vorkommen von Wörtern in Dokumenten, können sie als eine interessante Brückentechnologie zwischen eher an Inhalten interessierten Geistes- und Sozialwissenschaftlichen (Sub-)Disziplinen und der Linguistik gesehen werden. Insbesondere für Bindestrich-Linguistiken wie die Diskurs- (Spitzmüller & Warnke 2011), Sozio- oder Polito-Linguistik (Niehr 2014) ergeben sich vielversprechende Anknüpfungspunkte.

Quantitative Methoden spielen seit den späten 1960er Jahren eine zentrale Rolle in der Korpuslinguistik (Kučera & Francis 1967). Verfahren, welche über bloße Häufigkeitsstatistiken hinausgehen und vielmehr auf einem statistischen Modell der Wortverteilungen im Text basieren, finden sich erst seit den 1990er Jahren (Dunning 1993). Der Ansatz des *Topic Modelling* knüpft an derartige wahrscheinlichkeitstheoretische Betrachtungen an und ist ein Bayes'scher Ansatz. Das heißt, es wird auf der Grundlage bedingter Wahrscheinlichkeiten das Auftreten von Wortformen betrachtet, wobei für alle unbekannten Auftretenswahrscheinlichkeiten, die sogenannten Modellparameter, eine theoretische Vorannahme, die sogenannte A-priori-Verteilung, definiert wird, die ein angenommenes Vorwissen beziehungsweise Nicht-Wissen über die Parameter enkodiert. *Topic Modelling* stellt eine Weiterentwicklung des sogenannten *Probabilistic Latent Semantic Indexing* (PLSI) dar,[1] einem für die semantische Analyse von Texten angepasstem Verfahren zur Berechnung von latenten Variablen auf der Grundlage bedingter Wahrscheinlichkeiten (Hofmann 1999).

Wir beginnen mit einer kurzen Darstellung der grundlegenden Modellierungsidee von Topic-Modellen (Abschnitt 2). Dabei zeigen wir, wie die Wortformen in einem Dokument bzw. einer Dokumentkollektion verteilt sind und wie diese Verteilung für die Identifizierung von Topics genutzt werden kann. Hierauf aufbauend skizzieren wir die wesentlichen Aspekte des ersten Topic-Modells, der *Latent Dirichlet Allocation* (LDA; Abschnitt 3). Im folgenden Abschnitt stellen wir Beispielanwendungen von Topic-Modellen auf verschiedenen Textquellen wie literarische Texte, Nachrichten und Social Media-Daten vor (Abschnitt 4). Schließlich gehen wir auf einige wichtige Probleme und Best Practices bei der fachwissenschaftlichen Anwendung von Topic-Modellen ein (Abschnitt 5) und fassen abschließend die Möglichkeiten ihrer Nutzung in den Sprachwissenschaften zusammen (Abschnitt 6).

---

1 PLSI ist auch bekannt unter der Bezeichnung *Probabilistic Latent Semantic Analysis* (PLSA).

2 Intuition

Um die Grundidee des *Topic Modeling* nachvollziehen zu können, wollen wir nachfolgenden Text von Ludwig M. Eichinger betrachten:

> In analoger Weise wandte sich Maria Fernández-Villanueva der Frage zu, welche Stellung die Wortbildungslehre in einem Germanistikstudium in Spanien haben solle. Im normalen Curriculum der spanischen Germanistikstudiengänge habe die Wortbildungslehre einen festen Platz, sowohl als Element der grammatischen Teile linguistischer Einführungskurse, als auch in spezifischen Seminaren etwa zu lexikalischer Morphologie. Auch in den Sprachkursen würden Lesestrategien zur Auflösung von Wortbildungen vermittelt. Allerdings sollte man sich über die Art und Weise sowie über das Ziel dieser Vermittlung nochmals Gedanken machen. Da die Muttersprache der Studierenden Spanisch, Katalanisch, Galizisch oder Baskisch ist, sei die kontrastive Perspektive zwischen Muttersprache und Deutsch als Fremdsprache auszunutzen, nicht nur um Wortbildungsprozesse zu erfassen, sondern auch um Funktionen zu erkennen, die in einer Sprache durch Wortbildung, in der anderen vielleicht häufiger durch syntaktische Strukturen oder andere Mittel zum Ausdruck gebracht werden. Rezeptions- und Produktionsschwierigkeiten sollten zum Anlass genommen werden, um Unterstützungsmaterialien zu Lese-, Exzerpt- und Reformulierungsstrategien und zur Verfertigung von Begleitheften für Referate oder Hausarbeiten zu entwickeln, die funktional die Wortbildung ausnutzen. (Eichinger, Meliss, Dominguez Vasquez 2008: 354)

Offenbar handelt der Text von der Relevanz der Wortbildungslehre beim Lernen von Deutsch als Fremdsprache. Naheliegende Themenfelder, in unserer Terminologie also Topics, lassen sich dabei durch die Zusammenfassung von im Text vorkommenden Wortformen beispielsweise wie folgt beschreiben:

1. Aspekte der Wortbildungslehre: Wortbildung, Wortbildungslehre, lexikalische Morphologie, Wortbildungsprozesse, Funktionen, syntaktische Strukturen, Rezeptions- und Produktionsschwierigkeiten etc.
2. Länder und Sprachen: Spanien, spanisch, Muttersprache, Spanisch, Katalanisch, Galizisch, Baskisch, Deutch, Fremdsprache etc.
3. Sprachdidaktik: Curriculum, Germanistikstudium, Germanistikstudiengänge, Wortbildungslehre, linguistische Einführungskurse, spezifische Seminare, Sprachkurse, Lesestrategien, Vermittlung, kontrastive Perspektive, Rezeptions- und Produktionsschwierigkeiten, Referate, Hausarbeiten etc.

Die Grundidee bei der *Latent Dirichlet Allocation*, dem ersten von Blei, Ng & Jordan (2003) eingeführten Topic-Modell, setzt genau an dieser Beobachtung an. Die erste Modellannahme besteht darin, dass jedes Dokument durch eine kleine Untermenge von global verfügbaren Themen charakterisiert werden kann (in unserem Beispiel die drei über die Begriffslisten charakterisierten Topics), wobei jedes Thema wiederum durch eine ebenfalls kleine Untermenge

des Gesamtvokabulars beschrieben wird. Für die Modellierung wird zweitens angenommen, dass diese Themen einen zentralen, wenn auch versteckten, oder latenten, Parameter bei der Generierung eines Textes darstellen – die Wahrscheinlichkeit, in einem Text oder Textabschnitt eine bestimmte Wortform anzutreffen, hängt wesentlich davon ab, welches Thema vorab ausgewählt worden ist. Die dritte Annahme besagt, dass wir bei der Analyse eines Textes oder einer Textkollektion mit dem Verfahren des Topic Modeling den für die Textproduktion vorausgesetzten generativen Prozess umdrehen können und aus den vorhandenen Daten, d. h. den Wortformen, welche den Text konstituieren, die latenten Topics als den „besten" Parameter inferieren können, welcher die vorliegenden Daten am besten erklärt.

Die Grundlage des Algorithmus bildet also eine Menge artifizieller Topics. Jedes dieser Topics ist eine Wahrscheinlichkeitsverteilung über das vorhandene Vokabular und gibt an, wie wahrscheinlich ein Wort in diesem Topic ist. Jedes Dokument wird nun als eine Wahrscheinlichkeitsverteilung über eben diese Topics dargestellt, welche wiederum angibt, wie wahrscheinlich ein Topic für das aktuelle Dokument ist. Die generative Annahme der Dokumententstehung wählt pro Wortposition im Dokument zuerst ein Topic (proportional zu seiner Wahrscheinlichkeit im Dokument) und danach ein Wort aus dem gewählten Topic (ebenfalls proportional zu seiner Wahrscheinlichkeit im gewählten Topic). Ausgehend von dieser Annahme und den tatsächlich vorhandenen Wörtern in den Dokumenten können Rückschlüsse auf die Struktur der Wahrscheinlichkeitsverteilungen der Dokumente über Topics sowie der Topics über Wörter gezogen und diese approximiert werden. Das Ergebnis ist ein featuretransparentes Verfahren, welches in Aufgaben wie dem Dokument-Clustering, der Bestimmung von Dokument-Ähnlichkeit, der Dokument-Klassifikation oder der explorativen Suche nutzbringende Informationen beitragen kann.

3 Topic Modeling

Wir wollen diesen Modellierungsansatz nun konkretisieren. Grundlage der Textrepräsentation bildet das sogenannten *bag of words*-Modell, d. h. wir gehen davon aus, dass wir die Reihenfolge der Wortformen im Text ignorieren können. Jeder Textkorpus C enthält eine Menge von D Dokumenten, und jedes Dokument wiederum eine Menge von N_d Wörtern (Tokens). Die Gesamtzahl der Tokens eines Korpus bezeichnen wir mit N, das Vokabular des Korpus aller voneinander verschiedener Wortformen (Types) mit V. Jedes Dokument ist eine parametrisierte Repräsentation des Vokabulars bezogen auf die Frequenz der Types im Dokument.

Zur Modellierung der oben skizzierten bedingten Wahrscheinlichkeiten nehmen wir an, dass jedes Dokument eine Mischung von (latenten) Topics ist und jedes Topic eine (beobachtbare) Mischung aus Wortformen.

Notation

P(z)	ist eine Verteilung über Topics z in einem Dokument
P(w|z)	sind die Verteilungen über Wortformen w für Topics z
$P(z_i = j)$	ist die Wahrscheinlichkeit, dass für i-te Wortform Topic j gezogen wird
$P(w_i \mid z_i = j)$	ist Wahrscheinlichkeit von Wortform w_i im Topic j

Die Wahrscheinlichkeit, zu welchem Topic eine Wortform gehört, können wir nun als Produkt der bedingten Wahrscheinlichkeit $P(w_i \mid z_i = j)$ mal der Wahrscheinlichkeit des Topics im Korpus $P(z_i = j)$ berechnen:

$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) P(z_i = j)$$

Um zu beschreiben, welche Topics für ein Dokument bzw. welche Wörter für ein Topic wichtig (eigentlich: wahrscheinlich) sind, schreiben wir

$$\varphi^{(j)} = P(w \mid z = j) \text{ und } \theta^{(d)} = P(z).$$

Tatsächlich beschreiben die latenten Variablen φ und θ zwei Multinominalverteilungen bzw. Matrizen, nämlich die Zuordnung von Wörtern zu Topics *(*Wort-Topic Matrix φ*)* sowie die Zuordnung von Dokumenten zu Topics *(*Dokument-Topic Matrix θ*)*. Die Wort-Topic Matrix hat zwei Dimensionen, *K* und *V*, wobei *K* die Anzahl der Topics im Modell bezeichnet (eine Größe, die vom Nutzer vorher festgelegt werden muss) und *V* das Vokabular. Jeder Wert ϕ_{kw} bezeichnet somit die bedingte Wahrscheinlichkeit, mit der eine Wortform *w* aus *V* in einem Topic *k* aus *K* auftritt. Die Dimensionen der Dokument-Topic-Matrix θ sind *K* und *D*, wobei *K* wieder die Anzahl der Topics im Modell bezeichnet und *D* die Anzahl der Dokumente im Korpus. Jeder Wert θ_{dk} bezeichnet also die bedingte Wahrscheinlichkeit mit der Topic *k* aus *K* in einem Dokument *d* aus *D* auftritt.

Für die Berechnung eines Topic-Modells müssen die beiden Matrizen φ und θ geschätzt werden. Hierfür werden für die skizzierte Modellierung folgende Festlegungen getroffen:

1. Die Themenverteilung $\theta^{(d)}$ von Topics zu Dokumenten ist eine Dirichlet-Verteilung mit Hyperparameter α (die vom Nutzer vorab festgelegt werden muss).[2]
2. Die Zuordnung von Wortformen zu Topics $\varphi^{(j)}$ ist ebenfalls eine Dirichlet-Verteilung mit Hyperparameter β (die vom Nutzer vorab festgelegt werden muss).
3. Die latenten Variablen werden aus den beobachteten Wortformen im Text durch statistische Inferenz abgeleitet, indem ein generativer Prozess simuliert wird, der die tatsächlich beobachtete a posteriori Verteilung der Wortformen am besten approximiert. (Meist wird hierfür das sogenannte Gibbs Sampling verwendet).[3]
4. Um sinnvoll interpretierbare Ergebnisse zu erhalten, müssen für die LDA-Verteilungen θ und φ zwei konkurrierende Ziele austariert werden: Einerseits sollen jedem Dokument so wenig Topics wie möglich eine hohe Wahrscheinlichkeit haben, andererseits sollen aber auch in jedem Topic so wenig wie möglich Wörter mit hoher Wahrscheinlichkeit enthalten sein. Triviale Lösungen wären jeweils jedem Dokument nur ein Topic zuordnen, was aber das Erreichen des zweiten Zieles erschwert, oder umgekehrt jedem Topic nur ein Wort zuordnen, was aber das Erreichen des ersten Zieles erschwert. In der praktischen Anwendung müssen also beide Ziele im Auge behalten werden. Weiterhin ist zu bedenken, dass probabilistische graphische Modelle sich allgemein nicht analytisch berechnen lassen, weil die Komplexität der Evidenzberechnung exponentiell mit der Anzahl der Trainingspunkte (Beobachtungen) steigt. Die effiziente Parameterschätzung in Topic-Modellen stellt also für deren praktische Anwendung eine zentrale Herausforderung dar, bei der allerdings in letzter Zeit große Fortschritte gemacht worden sind (u. a. Schuster 2015; Teichmann 2016).
5. In den vergangenen Jahren wurden zahlreiche Varianten des ursprünglichen LDA-Modells weiterentwickelt, welche unterschiedliche Aspekte in die Modellierung von Textkollektionen mit einbeziehen. Das *Correlated Topic*

---

2 Die Dirichlet-Verteilung legt fest, wie wahrscheinlich eine multinomiale Verteilung ist. Betrachten wir z. B. einen Würfel mit 6 Augen, dann gibt die Multinomialverteilung an wie wahrscheinlich 1, 2 etc. auftreten. Die a priori Dirichlet-Verteilung sollte derjenigen Multinomialverteilung eine hohe Wahrscheinlichkeit geben, die allen Augenzahlen gleiche Wahrscheinlichkeit zuweist. Aber auch andere Verteilungen sind denkbar, etwa Würfel, bei denen nur wenige Augenzahlen eine hohe Wahrscheinlichkeit erhalten. Die Steuerung erfolgt über den Hyperparameter.

3 Beim Gibbs Sampling wird für jedes Wort abhängig von allen anderen Zuordnungen seine Topiczuordnung berechnet. Die hochdimensionale Verteilung wird durch wiederholtes Ziehen von niedrigdimensionalen Variablen simuliert. Von Verteilung über z ausgehend, werden also φ und θ iterativ approximiert.

Model beispielsweise (Blei & Lafferty 2006) berücksichtigt, dass bestimmte Themen bevorzugt gemeinsam in Dokumenten auftreten. Beim *Author-Topic-Modell* (Rosen-Zvi et al.2004) wird die Autoren-Präferenz für bestimmte Themen mit modelliert. Zeitdynamische Topic-Modelle (Jähnichen 2016) erlauben die Modellierung der Veränderung thematischer Zusammensetzungen in diachronen Korpora. Darüber hinaus existieren zahlreiche weitere Varianten, welche etwa die Verknüpfung von Dokumenten mit externen Klassenvariablen erlauben (*supervised LDA*) oder gemeinsame Themen in multilinguale, alignierten (Mimno et al. 2009) und nicht-aliginierte (Boyd-Graber & Blei 2009) Dokumenten über Sprachgrenzen hinweg finden können. Je nach Anwendungskontext und Forschungsfrage können diese zusätzlichen Aspekte wertvolle Informationen für eine Analyse mitliefern. In vielen Fällen können jedoch dieselben Auswertungen mit dem ursprünglichen LDA-Modell in Verbindung mit den Dokument-Metadaten in ähnlicher Qualität erstellt werden.

4 Anwendungsbeispiele

Topic-Modelle können grundsätzlich überall dort eingesetzt werden, wo sehr umfangreiche Textressourcen vorliegen, die im Zuge einer fachwissenschaftlichen Analyse nach inhaltlichen Kriterien strukturiert werden sollen, die möglichst direkt aus den Texten abgeleitet worden sind. Dabei sind die Anwendungsbereiche genau so vielfältig, wie die auszuwertenden Textressourcen und reichen von der Marktanalyse über den Sicherheitsbereich bis in die traditionellen Geistes- und Sozialwissenschaften, etwa der Literaturwissenschaft oder der Politikwissenschaft. Um die Bandbreite der Einsatzmöglichkeiten zu verdeutlichen, möchten wir im Folgenden Beispiele für den Einsatz von Topic-Modellen auf unterschiedlichen Textsorten und mit unterschiedlichen Erkenntnisinteressen vorstellen.

4.1 Literaturstudien

Bereits einleitend haben wir auf das *Culturomics*-Projekt sowie Morettis Forschungsprogramm eines *Distant Reading* in den Literaturwissenschaften hingewiesen. Tatsächlich haben Topic-Modelle in den letzten Jahren zu einer Reihe interessanter Forschungsprojekte beigetragen, bei denen große, diachrone Literatursammlungen auf textübergreifende Zusammenhänge und Strukturen untersucht wurden. Tangherlini & Leonhared (2013) beispielsweise greifen Morettis Formulierung des „Great Unread“, der großen Menge ungelesener

Texte in den klassischen Kanons der Einzelphilologien, direkt auf. Unter Nutzung von Topic-Modellen entwerfen sie einen Forschungsablauf, mit dem sie die vielen tausend dänischen Werke im *Google Books*-Projekt auf konkrete Fragestellungen hin näher untersuchen. So untersuchen sie unter anderem, wie sich die 1870 auf Dänisch erschienene Übersetzung von Darwins *Entstehung der Arten* in der dänischen Literatur niedergeschlagen hat. Technisch gelingt dies, indem ein Topic-Modell, berechnet auf Darwins Werk, auf Literaturtexte angewendet wird. So können Textpassagen identifiziert werden, die thematisch bzw. von ihrem Wortgebrauch her Ähnlichkeiten zu den damals revolutionären, naturwissenschaftlichen Betrachtungsweisen aufweisen. Mit diesem Ansatz können Tangherlini & Leonhard (2013) einen großen Einfluss der neuen, durch Darwin inspirierten Sicht auf den Menschen in der Literatur nachweisen. Gleichzeitig identifizieren sie bekannte, populäre Texte passend zu den Thematisierungsweisen, können aber auch heute bereits in Vergessenheit geratene Werke aus dem *Google Books*-Korpus zu Tage fördern. Mit ihrer Methode gelingt es also, eine umfassende Analyse des Einflusses von Darwins Werk auf die dänische Literatur zu zeichnen, die zudem reproduzierbar und intersubjektiv nachvollziehbar Gegenstand neuer literaturwissenschaftlicher Auseinandersetzungen werden kann. Mit Nachvollziehbarkeit und Güterkriterien bei der Anwendung von Topic-Modellen beschäftigen sich auch Jockers & Mimno (2013). Sie untersuchen in ihrer Arbeit 3.279 fiktionale Werke aus den Jahren 1899 bis 1970 mit einem $K = 500$ topics umfassenden LDA-Modell. Ihre Fragestellung lautete, wie sich externe Faktoren wie Geschlecht, Nationalität oder Geburtsjahr auf die Themenwahl von Autoren in der englischsprachigen Literatur des 19. Jahrhunderts auswirken. Methodisch entwickelten sie dafür ein Verfahren, mit dem sich Signifikanz der Themenabhängigkeit von solchen externen Faktoren bestimmen lässt.

4.2 Wissenschaftsgeschichte

Nicht nur große Literaturkorpora lassen sich mit Topic-Modellen untersuchen. Mittlerweile wird das Verfahren erfolgreich in der Wissenschaftsgeschichte eingesetzt, um Schwerpunkte der thematischen Entwicklungen einzelner Disziplinen über den Verlauf des 20. Jahrhunderts nachzeichnen zu können. Dazu werden (retro-)digitalisierte Archive von wissenschaftlichen Fachjournalen ausgewertet. Beye Riddel (2014) beispielsweise analysiert die Entwicklung von Themen in vier germanistischen Journalen, die über den Archivanbieter JSTOR zur Verfügung stehen. Sein Korpus umfasst mehr als 22.000 Fachartikel, die zwischen 1928 und 2006 veröffentlicht wurden. Hit Hilfe einer LDA-Analyse identifiziert er Trends abnehmender (etwa Sprachpädagogik), zunehmender

(etwa feministische Literatur) und wiederkehrender (z. B. Grimm'sche Märchen) Themen über die Zeit. Für einzelne Themen aus einzelnen Zeitabschnitten lassen sich repräsentative Fachartikel identifizieren, die eine tiefere inhaltliche Charakterisierung der Themen erlauben. Zudem können die Informationen über Topic-Zusammensetzungen jedes einzelnen Dokuments dazu genutzt werden, thematisch ähnlich zusammengesetzte Artikel zu finden. Eine ähnliche Analyse liefern Goldstone & Underwood (2012) mit ihrer Analyse der Artikel in PMLA, dem seit 1884 erscheinenden *Journal of the Modern Language Association of America*. Ihre Analyse ist besonders interessant, als dass beide Forscher unabhängig voneinander eine Topic-Modell-Analyse auf demselben Datensatz durchgeführt haben und beide Ergebnisse miteinander vergleichen. Dadurch zeigt sich das große Potenzial von solchen Analysen für intersubjektive Nachvollziehbarkeit, aber auch die möglichen Probleme und Fallstricke die auftreten können. Der Artikel gibt exzellente Hinweise darauf, wie Ergebnisse von Topic-Modellen in Abhängigkeit vom jeweiligen Forschungsinteresse interpretiert werden können.

4.3 Zeitgeschehen

Topic-Modelle eignen sich nicht nur dazu, in historischen Daten Zusammenhänge zu erkennen, sondern auch zur Beobachtung und Einordnung aktueller Ereignisse des Zeitgeschehens. Ein Beispiel liefern die vom Projekt *Deutscher Wortschatz* an der Universität Leipzig berechneten *Wörter des Tages*, welche eine automatische Schlüsselwortextraktion auf Artikeln der Tagespresse durchführen, wobei die thematischen Zusammenhänge der zugrundeliegenden Artikel anhand eines Topic-Modells berücksichtigt werden. Auf diese Weise entstehen themenspezifische Wortwolken, mit denen sich das Geschehen eines Tages zusammenfassen und auf einen kurzen Blick erfassen lässt. Abbildung 15.1 zeigt eine solche Zusammenfassung für den 20. Januar 2017, bei der die Amtseinführung des amerikanischen Präsidenten neben politischen Ereignissen in Gambia und verschiedenen Sport- und Boulevardereignissen sichtbar wird.

Neben den klassischen Zeitungsmedien können Topic-Modelle aber auch bei der Analyse sozialer Medien eingesetzt werden. So wurden im EU-Projekt *Slándáil*, das zum Gegenstand hatte, die Nutzung sozialer Medien im Katastrophenfall zu untersuchen, auch Nachrichten in Facebook und Twitter mit Topic-Modellen analysiert, die während der großen Flut in Sachsen 2013 von Betroffenen abgesetzt worden sind (Gründer-Fahrer et al. 2018). Die Ergebnisse zeigen zum einen, dass Facebook und Twitter von den Nutzern offenbar unterschiedlich genutzt werden (vgl. Abb. 15.2): Die Nutzereinträge in Facebook, das die Kommunikation von Nutzern in ihren sozialen Netzwerken zum Ziel hat und

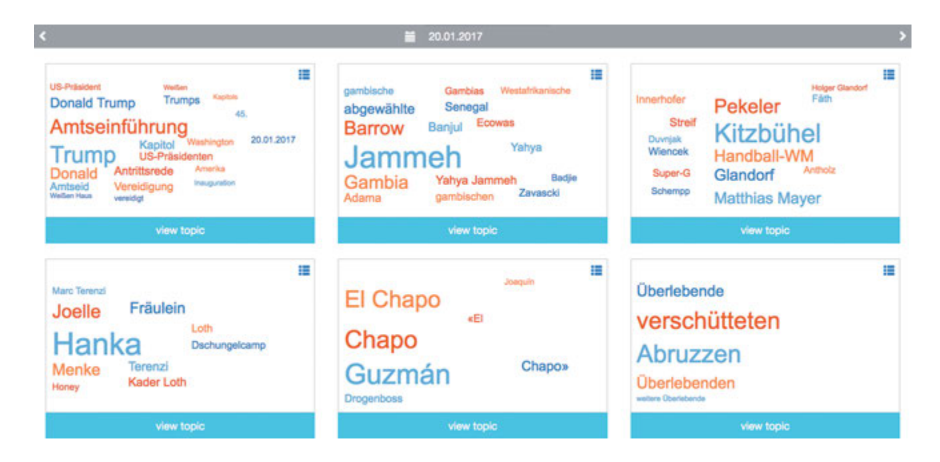


Abb. 15.1: Wörter des Tages vom 20. 1. 2017.[4]

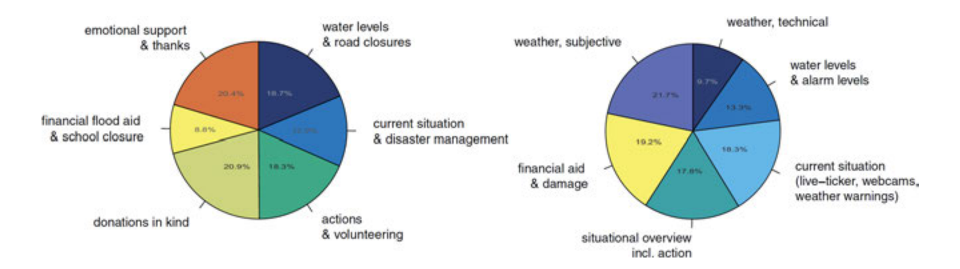


Abb. 15.2: Topics in Facebook (links) und Twitter (rechts).

längere Texte sowie emotionale Wertungen erlaubt, haben einen Schwerpunkt bei Topics, die sich der emotionalen Bewältigung der Flutkatastrophe sowie der Organisation von Hilfeleistungen zuordnen lassen. Die Nutzereinträge in Twitter, das auf die Verbreitung sogenannter Microblogs an ein Netzwerk aus Followern ausgerichtet ist, werden Schwerpunkte in der Vermittlung sachlicher bzw. subjektiver Information über Wetter und Pegelstände sowie Schadensmeldungen erkennbar.

Zum anderen spiegeln die Topics im zeitlichen Verlauf deutlich die verschiedenen Phasen der Flutkatastrophe wieder, wie es die vorstehende Abbildung 15.3 für Facebook verdeutlicht: Zunächst steht der Austausch über die tatsächliche Lage im Vordergrund, dann wird Hilfe organisiert und schließlich geht es um Spendenaufrufe, emotionale Aufarbeitung und gegenseitigen Dank.

4 http://wod.corpora.uni-leipzig.de/de/de/2017/01/20

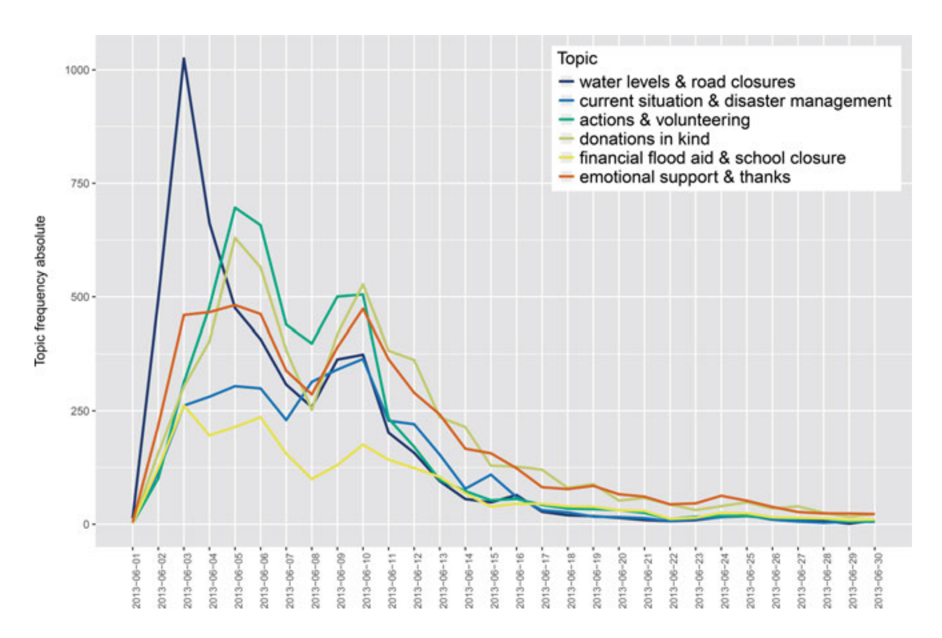


Abb. 15.3: Zeitlicher Verlauf von „Flut-Topics“ in Facebook.

5 Evaluierung und Best Practices

Zur Aufteilung einer Dokumentkollektion in Themen benötigt der LDA-Algorithmus drei sogenannte Hyperparameter: *K*, die Anzahl an Themen, die inferiert werden soll sowie α und β, die Dirichlet-Parameter, welche die Topic-Dokument- bzw. die Wort-Topic-Verteilung steuern. Die Wahl dieser Hyperparameter hat unmittelbare Auswirkungen die Qualität und Interpretierbarkeit der Modelle. Je nachdem, wie viele *K*-Themen gefunden werden sollen, lassen sich die Bedeutungen der einzelnen Topics unterschiedlich gut aus den in ihnen enthaltenen, hochwahrscheinlichen Begriffen rekonstruieren. Wird nur eine kleine Anzahl an Topics extrahiert, werden diese eher durch abstrakte, allgemein Begriffe geprägt. Ihnen lässt sich dann schlecht ein konkreter thematischer Sinn zuordnen. Wird ein sehr hohes *K* gewählt, so können Bedeutungen der Topics stark auf seltene Aspekte in der Dokumentkollektion bezogen sein. Die Anzahl an Themen sollte daher mit Bedacht gewählt und von verschiedenen forschungspraktischen Erwägungen abhängig gemacht werden. Dazu können mehrere Modelle mit unterschiedlichen *K* berechnet und evaluiert werden. Die Qualität eines Topic-Modells kann in diesem Zusammenhang mit verschiedenen Verfahren beurteilt werden (Maier et al. 2018). Qualitativ kann beurteilt werden, ob den einzelnen Topics (intersubjektiv) ein bestimmter Sinn zuge-

schrieben werden kann. Quantitativ kann gemessen werden, ob hochwahrscheinliche Begriffe eines Topics tatsächlich gemeinsam miteinander in Dokumenten auftreten (*Topic Coherence*). Diese Ansätze eignen sich gleichzeitig um geeignete Werte für die beiden Dirichlet-Parameter α und β zu finden. Hohe Werte (>= 1) dieser Parameter sorgen dafür, dass sich die Wahrscheinlichkeiten von Topics bzw. Worten gleichmäßiger über die einzelnen Dokumente bzw. Themen verteilen. Dadurch können Themen allgemeiner und weniger trennscharf werden. Kleine Werte für α und β (< 1) sorgen dagegen dafür, dass die Themen- bzw. Wortzusammensetzungen spezifischer werden, bei sehr kleinen Werten jedoch ebenfalls zu einer schlechten Modellqualität führen kann.

Eine weitere Schwierigkeit im Umgang mit Topic-Modellen ergibt sich aus deren eingeschränkter Reproduzierbarkeit aufgrund ihrer Berechnung mit stochastischen Inferenzverfahren. In aller Regel arbeiten computergestützte statistische Modelle und Verfahren für die automatische und semi-automatische Sprachanalyse robust, was heißt, dass eine wiederholte Messung gleiche Ergebnisse erzielt. Diese Eigenschaft erfüllen Topic-Modelle nur bedingt. Die Schätzmechanismen für die Parameter eines Topic-Modells, meist implementiert unter Verwendung von Gibbs Samplern oder variationellen Inferenzmethoden, basieren auf der iterativen Annäherung an einen Optimalzustand, also einen Zustand an dem das Modell die zugrunde liegenden Daten bestmöglich erklärt. Die Startwerte für die Parameter, von denen aus die Annäherung an das Optimum begonnen wird, werden im Normalfall zufällig initialisiert (engl. *seed*). Im Verlauf der Inferenz werden aus den geschätzten bedingten Wahrscheinlichkeiten Stichproben gezogen (engl. *sampling*) und die Parameter des Modells iterativ aktualisiert. Durch den Einfluss dieser Zufallsprozesse zusammen mit den Verteilungseigenschaften natürlicher Sprache in den Dokumenten, als auch in der gesamten Dokumentenkollektion (bedingt durchs Zipf'sche Gesetz), kann es dabei zu unterschiedlichen Resultaten, also Topic Verteilungen, kommen. Statt der analytisch optimalen Lösung der Modellgleichung (das globale Optimum, welches allerdings nicht berechenbar ist) finden die stochastischen Inferenzprozesse also nur lokale Optima innerhalb des Raums aller möglichen Wahrscheinlichkeitsverteilungen von φ und θ. Studien zur Reliabilität haben gezeigt, dass sich die Modelle, je nach Kollektion und Inhalt, hinsichtlich der Interpretierbarkeit und Ähnlichkeit der Topics lediglich zu ca. 50–80 % reproduzieren lassen (Maier et al. 2018).

Diese Einschränkung muss in Anwendungen, die auf eine inhaltliche Interpretation der Topics angewiesen sind, bedacht werden. Zudem wurden im Zuge der Untersuchung dieses Phänomens Strategien vorgeschlagen, um diesem Problem zu begegnen. Die einfachste Strategie besteht darin, wiederholte Berechnungen mit dem gleichen Startwert zu initialisieren. Dies führt zu sta-

bilen Ergebnissen, suggeriert jedoch nur eine vermeintliche Stabilität der Inferenz. Wesentlich vielversprechender sind Strategien, die eine vorherige Initialisierung der Themenzughörigkeiten mit Hilfe von vorgeschalteten Analysen vornehmen (Lancichinetti et al. 2015), die Stichproben während der Schätzung durch andere Wörter im Dokumentkontext zu beeinflussen (Koltcov, Koltsova & Nikolenko 2014), oder versuchen die Topics mit deduktiv festgelegten Wortkontexten aus einem bestimmten Domänenvokabular zu fixieren (Andrzejewski Zhu & Craven 2009).

6 Zusammenfassung

Topic-Modelle stellen ein mächtiges Werkzeug dar, um große Textmengen in einer Vielzahl von Anwendungskontexten zu erschließen. Insofern thematische Zusammenhänge modelliert werden, können sie für die philologische Forschung ganz direkt überall dort einen Beitrag leisten, wo die Auswertung von Inhalten von Interesse ist. Dies ist mit Sicherheit für die oben genannten Beispiele der Literaturstudien unter dem *Distant Reading*-Paradigma sowie für die Wissenschaftsgeschichte von großem Interesse. Aber auch sprachwissenschaftliche Arbeiten jenseits der konkreten Inhalts- bzw. Themenebene können in der Arbeit mit digitalen Textquellen von Topic-Modellen profitieren. Mit der Zurückdrängung des Szientismus in der Linguistik sowie in der Inhaltsanalyse als alleingültiges Paradigma (vgl. Fühlau 1981) ist die Notwendigkeit der Beachtung vielfältiger, darunter auch inhaltlich-thematischer, Kontexte bei der wissenschaftlichen Auseinandersetzung mit Sprachhandlungen offensichtlich geworden. Forschungsrichtungen wie die Begriffsgeschichte/Historische Semantik oder Vokabularanalysen in der Sozio- und Polito-Linguistik können in diesem Zusammenhang durchaus von einer thematischen Vorselektion ihres (digitalen) Ausgangsmaterials profitieren. Im Projekt *ePol – Postdemokratie und Neoliberalismus* (Wiedemann, Lemke & Niekler 2013) beispielsweise wurde die Frage untersucht, ob sich im öffentlichen Diskurs der Bundesrepublik Deutschland die Zunahme einer „Alternativlosigkeitsrhetorik“ in Bezug auf politische Begründungen beobachten lässt (Ritzi & Lemke 2015). Gesucht wurde nach Begriffen wie „alternativlos“, „unabdingbar“ oder „unverzichtbar“ in einem repräsentativen Zeitungskorpus von 1949 bis 2011. Zwar gibt schon eine einfache Frequenzanalyse für diese Begriffe in allen Artikeln des Politikressorts über die Zeit gewissen Anhaltspunkte für die Konjunktur von Alternativlosigkeit als Begründungsmuster. Richtig aussagekräftig werden solche Vokabularanalysen aber erst, wenn die Grundgesamtheiten, in denen die Frequenz-

messungen stattfinden nach bestimmten thematischen Kontexten vorgefiltert werden. Grenzt man das Untersuchungskorpus vor der Messung auf europapolitische, sicherheitspolitische oder gesundheitspolitische Themen ein, lassen sich ganz spezifische Beobachtungen für die Verbreitung des Begründungsmusters machen, sowie thematisch kohärente Beispieltexte für eine weitere qualitative Analyse filtern. Topic-Modelle sind in diesem Szenario nicht ein Endergebnis der Analyse selbst, sondern stellen lediglich ein exzellentes Vorverarbeitungswerkzeug dafür zur Verfügung. Im Sinne dieser Vielfalt von Verwendungsmöglichkeiten von Topic-Modellen stellen auch Goldstone & Underwood (2012) in ihrer vergleichend-experimentellen Studie fest: „A topic model doesn't just show you what people are writing about [...]. It can also show you how they're writing. [...] To put this another way, topic modeling can identify discourses as well as subject categories and embedded languages." (Goldstone & Underwood 2012) Dabei muss betont werden, dass der vollautomatische Ansatz keineswegs bedeutet, dass den sie nutzenden Forscherinnen und Forschern Interpretation und das Abwägen zwischen Alternativen abgenommen würde. Tatsächlich erfordert die Interpretation eines Topic-Modells und dessen Einbindung in ein kohärentes Forschungsdesign ein hohes Maß an Kreativität und den Einsatz von Kontext- bzw. Fachwissen der Forschenden, um aus den numerisch repräsentierten Clustern erhellende und zulässige Schlüsse zu ziehen. Dass die Suche nach geeigneten Modellparametern und die stochastischen Inferenzprozesse für wiederholte Modellberechnungen immer wieder zu leicht anderen Ergebnissen führen, muss nicht zwingend als ein methodisches Defizit betrachtet werden. Reproduzierbarkeit und Reliabilität an ein methodisches Verfahren, die aus einer szientistischen Perspektive als zwingende Bedingungen erscheinen, haben in einer poststrukturalistischen Perspektive deutlich geringeren Stellenwert. Hier kann die mangelnde Stabilität der Modellergebnisse sogar als Vorteil erscheinen. Goldstone & Underwood betonen den Fakt, dass leicht variierende Modelle aus wiederholten Modellberechnungen mit unterschiedlichen Parametern immer wieder neue Perspektiven auf das zugrunde liegende Textkorpus ermöglichen. Dabei können alle diese Perspektiven gleichsam Gültigkeit beanspruchen, insofern sie direkt aus den Daten abgeleitet sind: „But they're all pictures of the same evidence and are by definition compatible. Different models may support different interpretations of the evidence, but not interpretations that absolutely conflict." Für die Nutzung computergestützter Verfahren in der philologischen Forschung mit ihren vielfältigen Methoden, Schulen und Paradigmen ist dies eine kaum zu unterschätzende Erkenntnis. Computergestützte Verfahren der automatischen Sprachverarbeitung liefern uns eben nicht die eine objektive Wahrheit auf die über Sprache vermittelten Forschungsgegenstände, sondern eröffnen

uns stattdessen vielfältige, neue Perspektiven, mit denen der forschende Geist in die Lage versetzt wird, Erkenntnis aus den mittlerweile händisch nicht mehr überschaubaren Textmengen zu gewinnen.

Literatur

Andrzejewski, David, Xiaojin Zhu & Mark Craven (2009): Incorporating domain knowledge into topic modeling via Dirichlet Forest priors. In: *Proceedings of the 26th Annual International Conference on Machine Learning*, 25–32. New York: ACM (ICML '09). Online verfügbar unter http://doi.acm.org/10.1145/1553374.1553378 (letzter Zugriff: 6. 11. 2017).

Beye Riddell, Allen (2014): How to read 22,198 journal articles: Studying the history of German studies with topic models. In Matt Erlin & Lynn Tatlock (Hrsg.): *Distant readings: Topologies of German culture in the long nineteenth century: Boydell & Brewer*, 91–114. Online verfügbar unter http://www.jstor.org/stable/10.7722/j.ctt5vj848.7 (letzter Zugriff: 6. 11. 2017).

Blei, David M. (2012): Probabilistic topic models. Surveying a suite of algorithms that offer a solution to managing large document archives. *Communications of the ACM* 55 (4), 77–84.

Blei, David M. & John D Lafferty. (2006): Correlated topic models. In: *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 113–120. Pittsburgh: MIT Press.

Blei, David M., Andrew Y. Ng & Michael I. Jordan (2003): Latent Dirichlet allocation. *Journal of Machine Learning Research* 3, 993–1022. http://www.cs.princeton.edu/~blei/papers/BleiNgJordan2003.pdf (letzter Zugriff: 6. 11. 2017).

Boyd-Graber, Jordan & David M. Blei (2009): Multilingual topic models for unaligned text. In: *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 75–82. Arlington: AUAI Press (UAI '09). http://dl.acm.org/citation.cfm?id=1795114.1795124Online verfügbar unter (letzter Zugriff: 6. 11. 2017).

Crane, Gregory (2006): What do you do with a million books? *D-Lib Magazine* 12 (3). http://www.dlib.org/dlib/march06/crane/03crane.html (letzter Zugriff: 30. 11. 2017).

Dunning, Ted (1993): Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19 (1), 61–74. Online verfügbar unter http://dl.acm.org/citation.cfm?id=972450.972454 (letzter Zugriff: 6. 11. 2017).

Eichinger, Ludwig M., Meike Meliss & Maria José Dominguez Vasquez (Hrsg.) (2008): *Wortbildung heute. Tendenzen und Kontraste in der deutschen Gegenwartssprache* (= Studien zur Deutschen Sprache 44), 353–356. Tübingen: Narr.

Fühlau, Ingunde (1981): Inhaltsanalyse versus Linguistik. *Analyse und Kritik* 3 (1), 23–46.

Goldstone, Andrew & Ted Underwood (2012): *What can topic models of PMLA teach us about the history of literary scholarship?* Online verfügbar unter https://tedunderwood.com/2012/12/14/what-can-topic-models-of-pmla-teach-us-about-the-history-of-literary-scholarship/ (zuletzt aktualisiert am 14. 12. 2012, letzter Zugriff: 9. 4. 2017).

Gründer-Fahrer, Sabine, Antje Schlaf, Gregor Wiedemann & Gerhard Heyer (2018): Topics and topical phases in German social media communication during a disaster. *Natural Language Engineering* 24 (2), 221–264.

Hofmann, Thomas (1999): Probabilistic latent semantic indexing. In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 50–57. New York: ACM (SIGIR ’99).

Jähnichen, Patrick (2016): *Time Dynamic Topic Models*. Universität Leipzig. Online verfügbar unter http://nbn-resolving.de/urn:nbn:de:bsz:15-qucosa-200796 (letzter Zugriff: 10. 4. 2017).

Jockers, Matthew L. & David Mimno (2013): Significant themes in 19th-century literature. *Poetics* 41 (6), 750–769. doi:10.1016/j.poetic.2013.08.005.

Koltcov, Sergei, Olessia Koltsova & Sergey Nikolenko (2014): Latent Dirichlet allocation: Stability and applications to studies of user-generated content. In: *Proceedings of the 2014 ACM Conference on Web Science*, 161–165. New York: ACM (WebSci ’14). Online verfügbar unter http://doi.acm.org/10.1145/2615569.2615680 (letzter Zugriff: 6. 11. 2017).

Kučera, H. & W. N. Francis (1967): *Computational analysis of present-day American English*. Providence, Rl: Brown University Press.

Lancichinetti, Andrea, M. Irmak Sirer, Jane X. Wang, Daniel Acuna, Konrad Körding & A. Nunes Amaral (2015): High-reproducibility and high-accuracy method for automated topic classification. *Physical Review* X 5 (1), 11007. doi:10.1103/PhysRevX.5.011007.

Maier, Daniel, Annie Waldherr, Peter Miltner, Gregor Wiedemann, Andreas Niekler, Gerhard Heyer, Alexa Keinert, Barbara Pfetsch, Thomas Häussler, Ueli Reber, Hannah Schmid-Petri, Silke Adam (2017): Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. Communication Methods and Measures, online vrfügbar unter https://doi.org/10.1080/19312458.2018.1430754 (letzter Zugriff: 12. 4. 2018).

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva P. Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, et al. (2011): Quantitative analysis of culture using millions of digitized books. *Science* 331 (6014), 176–182. doi:10.1126/science.1199644.

Mimno, David, Hanna M. Wallach, Jason Naradowsky, David A. Smith & Andrew McCallum (2009): Polylingual topic models. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing:* Volume 2, 880–889. Stroudsburg: Association for Computational Linguistics (EMNLP ’09). http://dl.acm.org/citation.cfm?id=1699571.1699627 (letzter Zugriff: 6. 11. 2017).

Moretti, Franco (2007): *Graphs, maps, trees. Abstract models for literary history*. London, New York: Verso.

Niehr, Thomas (2014): *Einführung in die Politolinguistik. Gegenstände und Methoden*. 1. Aufl. Göttingen: Vandenhoeck & Ruprecht.

Ritzi, Claudia & Matthias Lemke (2015): Is there no alternative? The discursive formation of neoliberal power. *Cybernetics and Human Knowing* 22 (4), 55–78.

Rosen-Zvi, Michal, Thomas Griffiths, Mark Steyvers & Padhraic Smyth (2004): The author-topic model for authors and documents. In: *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*. Arlington: AUAI Press, 487–494. http://dl.acm.org/citation.cfm?id=1036843.1036902 (letzter Zugriff: 6. 11. 2017).

Schuster, Ingmar (2015): *Gradient importance sampling*. Cornell University Library. arXiv:1507.05781 (letzter Zugriff: 6. 11. 2017).

Spitzmüller, Jürgen & Ingo Warnke (2011): *Diskurslinguistik. Eine Einführung in Theorien und Methoden der transtextuellen Sprachanalyse*. Berlin, New York: de Gruyter.

Tangherlini, Timothy R. & Peter Leonard (2013): Trawling in the sea of the great unread. Subcorpus topic modeling and humanities research. *Poetics* 41 (6), 725–749. doi:10.1016/j.poetic.2013.08.002.
Teichmann, Christoph (2016): *Markov chain Monte Carlo sampling for dependency trees*. Dissertation, Fakultät für Mathematik und Informatik, Universität Leipzig 2016.
Wiedemann, Gregor, Matthias Lemke & Andreas Niekler (2013): Postdemokratie und Neoliberalismus – Zur Nutzung neoliberaler Argumentationen in der Bundesrepublik Deutschland 1949–2011. Ein Werkstattbericht. *Zeitschrift für politische Theorie* 4 (1), 99–116.

Register

https://doi.org/10.1515/9783110538663-017

Autorinnen und Autoren

Daniel Baumartz, Goethe-Universität Frankfurt, Robert-Mayer Straße 10,
D-60325 Frankfurt a. M., E-Mail: baumartz@stud.uni-frankfurt.de

Michael Beißwenger, Universität Duisburg-Essen, Institut für Germanistik, Berliner Platz 6–8,
D-45127 Essen, E-Mail: michael.beisswenger@uni-due.de

Mirjam Blümm, Niedersächsische Staats- und Universitätsbibliothek Göttingen,
Platz der Göttinger Sieben 1, D-37073 Göttingen, E-Mail: bluemm@sub.uni-goettingen.de

Matthias Boenig, Berlin-Brandenburgische Akademie der Wissenschaften, Jägerstraße 22/23,
D-10117 Berlin, E-Mail: dta@bbaw.de

Hans-Jürgen Bucher, Medienwissenschaft, Universität Trier, Universitätsring 15, D-54286 Trier,
E-Mail: bucher@uni-trier.de

Ruxandra Cosma, Universität Bukarest, Fakultät für Fremdsprachen,
Abteilung für Germanische Sprachen, Str. Pitar Mos 7–13, RO-010451 Bucuresti,
E-Mail: ruxandra.cosma@lls.unibuc.ro

Martine Dalmas, Sorbonne Université – Centre de linguistique en Sorbonne (EA 7332),
108, bd Malesherbes, F-75850 Paris Cedex 17, E-Mail: martine.dalmas@sorbonne-universite.fr

Kerstin Eckart, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart,
Pfaffenwaldring 5b, D-70569 Stuttgart, E-Mail: kerstin.eckart@ims.uni-stuttgart.de

Frank Fischer, National Research University Higher School of Economics, School of Linguistics,
Ul. Staraya Basmannaya 21/4, of. 207, RU-105066 Moskau, E-Mail: ffischer@hse.ru

Markus Gärtner, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart,
Pfaffenwaldring 5b, D-70569 Stuttgart, E-Mail: markus.gaertner@ims.uni-stuttgart.de

Alexander Geyken, Berlin-Brandenburgische Akademie der Wissenschaften,
Jägerstraße 22/23, D-10117 Berlin, E-Mail: dta@bbaw.de

Rüdiger Gleim, Goethe-Universität Frankfurt, Robert-Mayer Straße 10, D-60325 Frankfurt a. M.,
E-Mail: gleim@em.uni-frankfurt.de

Thomas Gloning, Institut für Germanistik, Justus-Liebig-Universität Gießen, D-35394 Gießen,
E-Mail: thomas.gloning@uni-giessen.de

Susanne Haaf, Berlin-Brandenburgische Akademie der Wissenschaften, Jägerstraße 22/23,
D-10117 Berlin, E-Mail: dta@bbaw.de

Wahed Hemati, Goethe-Universität Frankfurt, Robert-Mayer Straße 10, D-60325 Frankfurt a. M.,
E-Mail: hemati@em.uni-frankfurt.de

Gerhard Heyer, Universität Leipzig, Abteilung Automatische Sprachverarbeitung,
Augustusplatz 10, D-04109 Leipzig, E-Mail: heyer@informatik.uni-leipzig.de

Erhard Hinrichs, Seminar für Sprachwissenschaft, Eberhard Karls Universität Tübingen,
Wilhelmstr. 19, D-72074 Tübingen, E-Mail: erhard.hinrichs@uni-tuebingen.de

https://doi.org/10.1515/9783110538663-018

Wolfram Horstmann, Niedersächsische Staats- und Universitätsbibliothek Göttingen, Platz der Göttinger Sieben 1, D-37073 Göttingen, E-Mail: horstmann@sub.uni-goettingen.de

Bryan Jurish, Berlin-Brandenburgische Akademie der Wissenschaften, Jägerstraße 22/23, D-10117 Berlin, E-Mail: dta@bbaw.de

Hannah Kermes, Universität des Saarlandes, Fachrichtung Sprachwissenschaft und Sprachtechnologie, Campus, D-66123 Saarbrücken, E-Mail: h.kermes@mx.uni-saarland.de

Jonas Kuhn, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Pfaffenwaldring 5b, D-70569 Stuttgart, E-Mail: jonas.kuhn@ims.uni-stuttgart.de

Marc Kupietz, Institut für Deutsche Sprache, R5 6–13, D-68161 Mannheim, E-Mail: kupietz@ids-mannheim.de

Henning Lobin, Justus-Liebig-Universität, Institut für Germanistik, Otto-Behaghel-Str. 10 D, D-35394 Gießen, E-Mail: Henning.Lobin@germanistik.uni-giessen.de

Alexander Mehler, Goethe-Universität Frankfurt, Robert-Mayer Straße 10, D-60325 Frankfurt a. M., E-Mail: mehler@em.uni-frankfurt.de

Karlheinz Mörth, Österreichische Akademie der Wissenschaften, Sonnenfelsgasse 19, A-1010 Wien, E-Mail: karlheinz.morth@oeaw.ac.at

Andreas Niekler, Universität Leipzig, Abteilung Automatische Sprachverarbeitung, Augustusplatz 10, D-04109 Leipzig, E-Mail: aniekler@informatik.uni-leipzig.de

Philipp Niemann, Abteilung Wissenschaftskommunikation, Institut für Germanistik, Karlsruher Institut für Technologie (KIT), Kaiserstr. 12, D-76131 Karlsruhe, E-Mail: philipp.niemann@kit.edu

Andrea Rapp, Technische Universität Darmstadt, Institut für Sprach- und Literaturwissenschaft, Dolivostraße 15, D-64293 Darmstadt, E-Mail: rapp@linglit.tu-darmstadt.de

Stefan Schmunk, University of Applied Sciences Darmstadt, Max-Planck-Str. 2, D-64807 Dieburg, E-Mail: stefan.schmunk@h-da.de

Roman Schneider, Institut für Deutsche Sprache, R5 6–13, D-68161 Mannheim, E-Mail: schneider@ids-mannheim.de

Katrin Schweitzer, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Pfaffenwaldring 5b, D-70569 Stuttgart, E-Mail: katrin.schweitzer@ims.uni-stuttgart.de

C. M. Sperberg-McQueen, Black Mesa Technologies, 259 State Road 399, Espanola NM 87532, E-Mail: cmsmcq@blackmesatech.com

Elke Teich, Universität des Saarlandes, Fachrichtung Sprachwissenschaft und Sprachtechnologie, Campus, D-66123 Saarbrücken, E-Mail: E.teich@mx.uni-saarland.de

Christian Thomas, Berlin-Brandenburgische Akademie der Wissenschaften, Jägerstraße 22/23, D-10117 Berlin, E-Mail: dta@bbaw.de

Gregor Wiedemann, Universität Leipzig, Abteilung Automatische Sprachverarbeitung, Augustusplatz 10, D-04109 Leipzig, E-Mail: gregor.wiedemann@informatik.uni-leipzig.de

Frank Wiegand, Berlin-Brandenburgische Akademie der Wissenschaften, Jägerstraße 22/23, D-10117 Berlin, E-Mail: dta@bbaw.de

Tanja Wissik, Österreichische Akademie der Wissenschaften, Sonnenfelsgasse 19, A-1010 Wien, E-Mail: tanja.wissik@oeaw.ac.at

Andreas Witt, Universität zu Köln, Institut für Digital Humanities / Sprachliche Informationsverarbeitung & Institut für Deutsche Sprache, Mannheim,
E-Mail: andreas.witt@uni-koeln.de & witt@ids-mannheim.de